当代学术思潮译丛



人工智能哲学

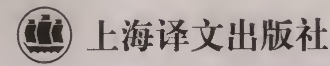
编者／[英] 玛格丽特·博登
译者／ 刘西瑞　王汉琦

上海译文出版社

当代学术思潮译丛

人工智能哲学

编者／[英] 玛格丽特·博登
译者／刘西瑞　王汉琦



图书在版编目(CIP)数据

人工智能哲学/(英)博登(Boden, M. A.)编著;刘西瑞,王汉琦译.—上海:上海译文出版社,2001.11
(当代学术思潮译丛)
书名原文:The Philosophy of Artificial Intelligence
ISBN 7-5327-2583-9

Ⅰ.人... Ⅱ.①博...②刘...③王... Ⅲ.人工智能-技术哲学-文集 Ⅳ.TP18-53

中国版本图书馆 CIP 数据核字(2001)第 03870 号

Margaret A. Boden
THE PHILOSOPHY OF ARTIFICIAL INTELLIGENCE
Oxford University Press, 1990
根据英国牛津大学出版社 1990 年版译出

人工智能哲学
[英]玛格丽特·博登 编著
刘西瑞 王汉琦 译

世纪出版集团
上海译文出版社出版、发行
上海福建中路 193 号
全国新华书店经销
上海汇杨印刷厂印刷

开本 850×1168 1/32 印张 19 插页 2 字数 407,000
2001 年 11 月第 1 版 2001 年 11 月第 1 次印刷
印数:0,001-5,100 册
ISBN 7-5327-2583-9/B·116
定价:23.60 元

译者的话

《人工智能哲学》(Philosophy of Artificial Intelligence)的原著是牛津大学出版社1990年出版的,由15篇文章组成。文章作者多是人工智能(AI)思想界的著名人物,而这15篇文章在AI发展史上则具有里程碑式的地位。文章写作时间起于1950年,止于1990年,它们是AI思想近半个世纪发展历程的记录。

原著主编兼作者玛格丽特·博登(Margaret Ann Boden)是英国科学院院士,英国苏塞克斯大学认知科学学院创始人兼院长,历任英国心理学会哲学、历史分会主席,以及英国皇家哲学学会理事,著有《人工智能史》一书。作为该领域的权威人士,她独具慧眼地从浩瀚的文献中挑选出这些代表性作品,并为本书撰写了长达20多页的导言。导言对书中每篇文章一一作了介绍,前后贯穿,对于读者完整地了解本书的全貌大有裨益。

作为论文集,一般来说不如专著那样系统和面

面俱到，但是换一个角度看，书中的文章多是作者原创性研究的代表之作，这些文章中蕴含着一个个新思想破土而出之际特有的创造性力度，有的立意深远，有的演绎缜密，不乏大家手笔，这种特色是一本综述性专著所无法比拟的。书中最早的文章是图灵写于1950年的“计算机器与智能”(第2章)。这是一篇公认的划时代之作，它提出了图灵机的理论模型，为现代计算机的出现奠定了理论基础。但是这篇文章的价值不仅仅体现在计算机领域，文中还提出了著名的图灵准则，在AI思想界，“图灵检验”已成为最重要的智能机标准，这几乎也是关于智能存在与否的唯一可操作的标准。这篇文章涉及的哲学理论问题还不止于此，具有无限长纸带的理想数字计算机与大脑的某些活动在本质上有什么共同之处，是同样值得深究的。这是一篇计算机界尽人皆知的文章，但是读过其全文的人并不多，今天读它，我们丝毫没有过时之感。在AI史上可与图灵检验相提并论的另一个标准是哲学家塞尔提出的，这是一个有趣的思想实验：塞尔被关在一间充满中文字条的屋子里，通过在窗口传递中文字条与外界发生联系，并靠一本英文指令书将各种中文字条配对，这样，他就可以正确地回答屋外中国人的提问，所以屋

外的中国人认为他是懂得中文的，可是，塞尔真的懂得中文吗？塞尔本人的回答是否定的。显然，图灵准则是行为主义的，而塞尔的思想实验却采取了相反的立场，强调意识的作用。塞尔在文章中列举了来自各个方面的对这一思想实验的反诘，然而这些反驳并没有动摇思想实验的根据：理解的实质决不是程式化操作。书中各篇文章都有其精彩之处，它们自成一体，又相互联系，向我们展示出 AI 研究领域的各个重要方面，以语境迥异的各家之言把我们带入 AI 思想领域的理论深层和研究前沿。

在科学家族中，没有一门学科比 AI 与哲学的关系更密切。科学从哲学母体中分离而出之后，仅在认识论层次上与哲学保有联系，然而 AI 却在其学科内部与哲学难解难分，难怪斯坦福大学的计算机教授要为计算机系的学生们讲授海德格尔（见第 13 章）。数理逻辑与计算机理论的关系自不待言，但维特根斯坦、胡塞尔和海德格尔的名字出现在 AI 问题的讨论中，也许会使一些读者感到意外。读一读德雷福斯的文章，我们就会看到，人作为“符号动物”，一切思想都必须由语言来表达，在此前提下，AI 发展中遇到的许多问题，哲学大师们都已经探讨过、预见过。诚如书中所言，许多没有出路的 AI 研究，“只

是因为对哲学家昔日的失败一无所知,才得以维持”。80年代,日本计算机界曾雄心勃勃地推出“第五代计算机”研制计划,结果是以“光荣失败”而告结束。所谓失败,是因为它没有实现当初的宏伟目标——以非数字化方式在日常范围内全面地模仿人类行为;所谓光荣,则是它得出了许多可资后人汲取的宝贵经验,其中之一就是AI必须同哲学联手。

与哲学联手,谈何容易!我们看看那些从AI诞生之日起就与之有不解之缘的问题:计算机或机器人是否能够成为认识的主体?机器的形式化过程与人的思维在本质上是相同的吗?如果承认计算机有智慧,这种智慧能够超过人类吗?……几十年过去了,人们仍然在讨论着类似的问题。这些问题之所以久久得不出答案,正是因为这些问题不能由计算机科学单独来回答,而要与哲学共同回答,可是20世纪的哲学家们远没有他们的先辈们那样幸运,爱因斯坦感叹过的那种每一个数学分支都会耗尽一个人短暂一生的情况,在20世纪的科学发展中有增无减,哪一位哲学家还能像他们的前辈笛卡尔、莱布尼兹那样身兼科学大师和哲学大师的双重身份呢?然而不入其境,焉能得其精髓,我们看到的是这样的情况:计算机科学家无暇顾及技术以外的问题,而哲学

家又只能在这个庞然大物的外围徘徊。当然，也有一些例外的情况，或是计算机专家对哲学问题发生了兴趣，或是哲学家达到了能与计算机专家对话的程度，而这一批杰出人物在世界范围内从最前面数起的话，他们的名字有许多都出现在这本书里了。

众所周知，关于 AI 的发展前景，有两个激烈对立的派别：乐观主义派别认为 AI 前途无量，AI 与人类智能并无实质性差别，它不仅能替代、而且将超过人类智能；悲观主义派别则把 AI 贬低为现代“炼金术”，认为它不过是零敲碎打和拼拼凑凑，它仅是在局部表面上对人类的摹仿，远未达到对人类智慧的实质了解。这种争论从一定意义上说已成为历史，因为 AI 的发展似乎并没有单独支持哪一派的立场，它既朝着深入认识、反映人类智慧的方向一步步前进，又未能实现当初的许多既定目标。争论双方都变得冷静了，意识到了对方观点中的合理因素，但是这并不意味着派别之争已经消失，我们在书中见到一对名词：强 AI 和弱 AI，这种提法比起以前的相互攻击显然温和了许多，但是它仍代表了两种截然有别的立场。强 AI 指出：AI 不仅仅是实现人类智能的工具，事实上它就等同于人类智能；弱 AI 则认为：AI 仅仅是实现人类智能的工具，对它的评价不应越

出这一范围。书中文章多少都带有两个派别的印迹,有些文章就是出自两派领袖人物之手,将它们对照起来看,更容易从中评判其得失与正误。除派别之分外,书中文章还分属两种不同的语境:技术的和哲学的。技术专家和哲学家在自己的范围里可以自圆其说,但这两种语境缺乏可通约性,只能是自话自说。把它们放在一起,同时从这两个角度认识AI,或许会形成思想碰撞,引发思想火花,促成某种新的融合。值得说明的是,这里的技术性文章并无艰深繁冗的理论推导,而着重于对推理思路的概述,不会给非计算机专业的读者造成理解上的困难。同样,哲学家的文章也着重于对基本思路的阐述,容易为哲学专业以外的人所接受,当然,对哲学研究者来说,可以悟出其中更深层的意义。

在AI内部又存在两大分支:一是传统AI——以符号逻辑为基础的算法系统,是由图灵、冯·诺伊曼规定出的那一套方法;另一个则是联结论——建立在统计分布规律之上的并行分布式系统,包括对大脑神经网络的模拟。联结论在很大程度上弥补了传统AI的不足,具有容错能力和较强的学习功能。本书的第1、11、12、14、15章都是涉及联结论的探讨。

1997年，在计算机诞生50年之际，计算机深蓝以2胜1负3平的成绩战胜了国际象棋世界冠军卡斯帕罗夫。人工智能，也就是计算机智能，再一次成为令世人关注的话题，它真的是所向无敌吗？它的能力究竟有没有界限？我们早已不怀疑计算机是迄今为止人类发明的最重要的机器，这不仅因为它深刻地改变了我们的生活，带来生产力的革命，影响到社会生活的每个角落，更因为它直接指向人类的本质特征——智慧。从一定意义上说，计算机科学，或更确切地说是AI，与物理、化学这些自然科学有着根本性的区别。AI虽然还带着年轻学科难以避免的简单化、局部性和缺乏统一理论的构造特点，但是它显然不同于研究那个“自在”世界的一般自然科学，它的对象是大脑思维活动，至于思维，我们用“意识、精神、主观”这样一些概念把它同“存在、物质、客观”区别开来。这是一个人类专有的领域，机器如何能涉足呢？在有些人看来，问题似乎很简单：思维是大脑这台生理机器的产物，既然构成大脑元件的物质与其他物质并无本质的不同，为什么大脑的活动不能由其他物质元件（机器）替代呢？然而，事实并非如此简单，对此还须作进一步追问：如果将大脑看作生理机器，那么这台机器是如何运作的呢？

它又如何产生出灵活多变的思想，而联想、幻觉、顿悟以及思想传递又是如何发生的？在不能具体解释其运作机理之前，这种生理机器的论点只能看作是一种猜想，而阐明这一猜想的根据还远未达到令人信服的程度。AI 的出现使我们朝作出解释的方向迈出了一步，至少在行为主义的标准下，机器可以再现智慧。（这也恰恰表明了行为主义的局限。）深蓝打败卡斯帕罗夫，关键是有象棋专家为其出谋划策，而象棋专家是以何种方式介入的，是经验的汇集——像专家系统那样，还是有理论上的新招，这一点尚未披露，但若是前者，同时伴以机器容量的增加和速度的提高，那么这令举世震惊的一步其实并没有什么更深刻的内容。人类智慧的那颗内核究竟藏于何处？AI 的功绩在于帮助我们一层一层剥去外皮，渐渐逼近那个令人神往的智慧之核，这也是 AI 对哲学的重大贡献。仔细读过这本书，我们将会从中看到更多的东西。

书中常常出现两个与 AI 特征相关联的概念：形式化和意向性。这两个概念在建立 AI 与哲学的联系上起着举足轻重的作用，这里作一简要介绍。

形式化(formalization)。形式化有狭义、广义两层意思。广义地说，一切被感知的事物，和由它们组

合而成的复杂事物，以及被意识到的精神活动，都以符号和其他某种形式在意识中形成对应物（图形、声音等代号形式），亦即构成意识的基本材料，这些东西总是表现为一定形式的，我们的思维活动就是通过对这种素材的组织而完成的。在表现为符号的（形式的）层次之下，有非形式的层次，也对应于某种脑的活动，但那是意识之外的东西，而处在意识之外，则不能直接为主体所把握，或者说，任何一个事物进入意识，都要借助于某种形式化过程。这里可以参照哲学家所说的语言的界限即为世界的界限而提出：形式化的界限即为思维的界限。狭义的形式化是由逻辑学规定的。本书中的形式化概念则常常仅对计算机和AI而言，是指一种事先规定好的运行方式。将某一过程形式化，也就是建立一种算法，将这一过程描述出来。任一事物，只要能够形式化，就可以由计算机来完成，其逆反推论也成立：任何不能形式化的事物，计算机都无法实现，所以我们又可以说，（狭义）形式化的界限就是计算机的界限。当然，由于联结论的出现，AI中的形式化要求有所放宽。但是即使在联结论中也必须以形式化为开始，才能建立分布式表述基础上的算法。这样，计算机能做什么和不能做什么的问题，就转化为形式化界

限在哪里的问题。狭义形式化的界限显然小于语言,而广义形式化的界限显然又大于语言。狭义形式化实际上是以固定形式构成的结构关系,而广义形式化则比较庞杂,包括与个别对象相对应的形式,也包括形式之间的关系结构。不难看出,实现形式化是 AI 的关键问题。

意向性(intentionality)。意向性被看作是区分个体的人和机器的根本特征之一:机器和人可以做同样的事情,但是人有意向性,而机器没有。那么,什么是意向性呢?简单说,意向性就是意识的指向性。在多数情况下,人的一举一动,一言一行,都是在意识的引导下完成的。而机器所做的任何事情都只不过是一种机械运行过程,这种过程是人事先指定好的,而不是机器自发产生的。那么主动性能否作为人和机器的本质区别之一呢?从一定意义上说是可以的。没有原发动机,无论完成多么复杂的任务,机器永远只能处在工具的地位上。无论多么能干的机器人,甚至成功地做出人所未曾预料到的事情,也仍然不能改变这一事实:它完成运作的先决条件——始发动力,必须由人从外部输入。人们常常历数计算机的优点:不知疲劳、不为感情所动、不会出现粗心的错误、不受外界干扰——而这一切恰恰是不具

有意向性而带有的特征,心绪不佳、感情冲动、注意力分散正是意向性的产物。所以要回答机器是否能完全和人一样地做某些事情,首先要回答的是:机器是否能够具有意向性。

有人对人的动机形成机制进行归纳,然后把它赋予机器(见第10章),这样的机器人自然更像一个真正的人,可是动机是与个体的独立意识相联系的,离开主体何谈动机,那么机器能获得主体意识吗?简单作出肯定或否定的回答,都缺乏足够的论据,这有待于我们对大脑构造进一步了解和对主体性作出进一步解释和界定,或者说AI的提问促使哲学家们从另一角度重新审视这一问题。

计算机可以具备许多与人类相似的能力:它可以记忆、推理、学习,凡是人类已经完成的事情,理想地看,计算机都能做到,但是由于它没有创造性,它就永远只能处在被动的地位上。意向性,以及与此相关的创造性,可以看作主动和被动的分界线。

以上对这两个概念的探讨是译者个人的理解,写在这里,供读者参考,以期引起进一步的讨论。

书中涉及到的深层理论问题还有许多,它会把我们带向AI技术背后的理性思考,使我们从一般的泛泛而谈进入涉及实质性的理论探讨。能将此书翻

译介绍给读者,我们感到十分高兴。

美国夏威夷大学教授泰尔斯(Tiles)女士惠赠本书的英语原著,泰尔斯教授曾在1990年中英哲学暑期班讲学,赠书乃是这一教事的延续,对她的美意当在此申谢。

本书得以出版还要感谢上海译文出版社的积极促成,他们注重学术价值,为了将此书译介给读者,不惜出资购买版权,并为本书的编辑出版付出诸多心血。

本书在翻译过程中得到许多师长、朋友的帮助。这里首先要感谢的是上海交通大学陈以鸿先生。陈先生在校阅过程中通读全部译稿,其负责精神十分感人,陈先生精湛的英语水平使译稿的质量大为提高。同时也要深深感谢西安交通大学的黄上恒副教授及许多未在这里提名的学者朋友们。

译稿虽经几次修改,但仍存在许多不妥之处,恳请读者不吝赐教。

刘西瑞　王汉琦

2000年8月于汕头大学医学院

目次

导言

人工智能(AI)有时被定义为：研究怎样制造计算机，并(或)为其编程，使其能做心灵所能做的那些事情。这些事情中有一些被公认为是需要智能的：开药方和(或)作医嘱，提供法律或科学咨询，证明逻辑或数学定理。另外一些事情则不同，它们与教育背景无关，是所有正常的成年人都能做到的(有时甚至人类以外的动物也能做到)，其特点是不受意识支配，如看到阳光下的物体和影子，找到穿过复杂地形的小路，把木桩塞进洞里，用母语讲话，以及运用自己的常识。

由于上述定义涵盖了与这两类心理能力有关的 AI 研究，所以它胜过把 AI 说成是让计算机去做“人类需要运用智能才能做的事情”的定义。然而它有一个预设假定：计算机所能做的就是人脑所能做的，计算机真的可以开处方，提建议，做推理，善理解。如果将 AI 定义代之以 AI 是“计算机的发展，而这些计算机的外在性能具有我们认为是属于人类心理过程的那些特征”，我们就有可能回避这一尚存争议的假定(同时也避开了计算机在做这些事情时是否采用了与我们

相同的方式这一问题)。这一适度的特征描述可以为一些AI工作者所接受,尤其是那些把眼光投向为商业目的而生产技术工具的人们。

但是也有不少人偏爱一个更有争议的定义,即把AI看作是**一般性的智能科学**,或更确切地说,看作是认知科学的智力内核。这样,它的目标就是提供一个系统的理论,该理论既可以解释(也许还能使我们复制)意向性的一般范畴,也可以解释以此为基础的各种不同的心理能力。它不仅要包括地球上各种生物的心理,而且还要包括全部可能存在的心灵。它必须告诉我们,智能是仅仅体现于那些具有大脑般的基本构造(包括由关联细胞组成的网络中的并行处理过程)的系统之中,还是也可以用某种别的方式来实现。这样,由于"计算机"已退出了这一定义,它们与这样一门科学的特殊关系必须加以确证。这一雄心勃勃的事业是否能够成功(如果能够,又怎样成功),抑或它是根本错误的想法,这个问题引发出许多与AI相关联的哲学问题。

因此,AI哲学(这里把AI看作是一般性的智能科学)同心灵哲学、语言哲学以及认识论紧密相联,同时又是认知科学哲学,特别是计算心理哲学的核心。计算心理学家们共同约定了四个哲学假定。对待心灵和智能,他们采用机能主义的方式,认为心理过程是能够被精确说明的过程,而心理状态则取决于它们与感觉输入、动作行为以及其他心理状态的因果关系。他们把心理学看成是对心理表象所藉以构成、解释和变换的那些计算过程的研究。他们把大脑视为一种计算系统,所关心的是它体现出何种函数关系,而不是哪些脑细胞体现出这些关系,或大脑生理机能怎样使这种体现方式成为可

能。他们虽然并不认为(AI 工作者也是如此)哪一些 AI 概念和计算机模型的方法论在认识智能方面可能是最有帮助的,但是他们都认为,某种 AI 概念一定会成为心理学理论基本内容的组成部分。

以与 AI 概念极为类似的概念对智能加以解释,是哲学家们长久以来的梦想,可以认为从柏拉图开始就是如此(见第 13 章)。在过去的许多个世纪中,这个梦想孕育了形而上学理论,产生出对心理机能的形式说明,甚至是解释模型——我们的脑海中会浮现出霍布斯、莱布尼兹和巴比奇这些人物。到了 20 世纪,这一思想的智力资源因三个方面的进展而更加丰富:形式计算理论,为实现形式上规定的计算而设计的功能计算机,以及神经元的发现。

这三个发展奠定了整个 AI 的基础,虽然有些研究对其中某一方面的发展利用得比较明显。当前的 AI 研究一般分为两大类别:"传统"类[或 GOFAI[①],即"有效的老式 AI"(Haugeland 1985)]和"联结论"。虽然它们之间的理论关系尚有争议,但是它们的历史关系是清楚的:它们是从同一个根上生长出来的分支,共同发轫于由神经心理学家兼精神病学家 W·麦卡洛克和数学家 W·皮茨合著的开创性之作(见第 1 章)。

麦卡洛克和皮茨合著的文章题为"神经活动内在概念的逻辑演算",这一题目本身就表明了传统 AI 和联结论 AI 的共同继承权。他们关于实施"逻辑演算"的设想,影响到冯·诺伊曼对数字计算机的设计,同时鼓舞着 AI 的先驱者们尝试建立

① GOFAI 是 Good Old Fashioned AI 的缩写。——译者

思维的形式模型。他们对于"神经活动"的讨论使赫布的细胞组合生理心理学理论获益匪浅,并促成了多种多样的神经网络模型——这正是今日联结论系统的先驱。

这篇文章之所以具有重大影响,主要是因为它虽有臆想的成份,但决不仅仅是一种推测。毋庸讳言,文章作者关于目的、学习、精神病学的神经体现方式——更不必说认识论、实在论、普遍性、价值以及数字的神经体现方式(McCulloch 1965)——的大胆构想只是提纲挈领式的论述。但是麦卡洛克和皮茨并不是简单地持有一般的唯物主义的立场,认为智能是由大脑实现的,他们**证明了**:一定类型的(可严格定义的)神经网络,原则上能够计算一定类型的逻辑函数。

他们知道,神经系统是由相互联系的细胞组成的,这些细胞的激活表现为全或无的形式,并取决于阈值和其他细胞的活动性。他们也了解图灵关于可计算数字的文章(Turing 1936),以及罗素和怀特海在命题演算方面的工作。他们在整合这些不同资料的基础上,证明了有关理想化神经网络逻辑特性的各种定理。例如,每个命题演算函数都可以由某种(类型相当简单的)网络来实现;每个网络计算出一个可由一台图灵机计算的函数;同时每个图灵可计算函数可以通过某个网络来计算。图灵机具有一条无限长的纸带,也就是说,它是一种数学上的理想形式,而不是一台实际的机器。由于神经网络是有限的,我们不可能通过证明一种一般化的、也许是无法实现的可能性,来恰当地解释被体现的智能。相反,我们必须确定哪些网络能够实现一些特殊的功能。这样,理论心理学的任务就是要设计出具有计算能力的网络,而这种计算原先是由心灵完成的。

AI的任务就是确定和设计这种网络,并通过构造工作模型获得补充,而作为特例的人类心理也包括在这项(关注实际的和可能的心灵的)任务之中。如果把"网络"看作是真实的神经联结的近似形式,那么我们就得到一个广泛的联结论的研究纲领。由于被解释成神经活动高度抽象的理想形式,我们的主要着眼点是二值逻辑,而不是真正的细胞联结性和阈值,所以典型的传统AI是以数字式信息加工方式出现的。麦卡洛克和皮茨此文的研究成果对于这两种类型的AI研究都具有开创性意义。60年代后期,神经网络研究一度低落,部分原因是将只适用于一小类网络的批评作了过分的普遍化(Minsky and Papert 1969)。本书后面第11-13章所讨论的内容即是神经网络研究再度兴起后的成果,它(与AI和心理学中大多数联结论模型一样)并没有试图从构造上与可识别的神经联结相对映,这一艰巨任务常常是由神经科学家来承担的(见第14章)。

图灵关于可计算数字的文章(Turing 1936),对两种AI研究的途径来说,都堪称理论上的奠基之作,文中将计算定义为:应用形式规则,对(未加解释的)符号进行形式操作。"有效过程"——一种可严格定义的计算过程——的一般观念,是通过数学演算的例子来说明的,但是(正如麦卡洛克和皮茨所意识到的那样),这就意味着,如果智能可以普遍地用在大脑中实现的有效过程来解释,那么一台普适的图灵机,或是某种与之近似的实际机器,就可以对其进行模拟。1950年时,图灵,还有其他人,已经制造出通用数字计算机,他们被用来模拟智能的某些方面。在"计算机器与智能"(第2章)一文中,他特地提出了这种机器能否思维的问题。

他指出，对这个问题的回答，不应当依据预设的（很可能是武断的）“思维”定义，而是应当问一问某种可构想出的计算机是否能够表演“模仿游戏”，才能作出判断。无论是做加法还是阅读十四行诗，一台计算机能以无法与人类回答相区别的方式来回答提问者的问题吗？这个问题（常常表达为一台计算机是否能通过“图灵检验”）包括三个方面：某个未来的计算机真的有能力以所设想的方式回答问题吗？无论在人类还是在计算机中，有效过程原则上能够生成这种性能吗？这种性能足以使计算机具备智能属性吗？图灵本人对每一问题的回答都是：“是的。”

图灵的立场受到来自三个迥然不同方面的攻击（这三个方面既不相互排斥，也没有必然的联系），每一方面都存在许多重要的变化形式。

第一类攻击采用了一套常见的反行为主义的论据，拒绝作为智能充分判据的模仿游戏——此外，并未拿出任何专门与AI有关的东西。然而，即使如反行为主义者所坚持的那样，意识体验对于智力来说是一个必要条件，如果不能在未加论证的直觉以外提供新的论据，以说明意识显然不可能产生于计算机的话，就不能证明智能计算机是不可能的，一般说来，反行为主义的论据所能证明的，充其量不过是：一台高性能的计算机并不需要是智能的。

支持AI的人会表示赞同，因为他们采取了机能主义的立场，认为智能必须包含某种系统式的因果过程（计算）。然而，行为，无论在表面上给人以何等深刻的印象，仍是来自某个庞大的事先存储的查寻表，而不同于结构式的（有可能反映出习俗心理的精神范畴的）过程和表象，因此行为不能看作是智能

的(Sloman 1986)。从图灵的判据来看,他只规定了,作为智能行为基础的原因是某种有效过程。而且,由于他没有明确指出,思维必须包括思想者内部的基本原因,他的判据就不排除通过魔术或偶然地引起的行为:果实从被风吹动的树上落下,掉到电传机键盘上,可能会"愚弄"一个正在玩模仿游戏的提问者。

对图灵立场的第二类攻击,集结着这样或那样主张计算机不可能有智能的另一些看法。有一种看法认为,图灵依据于言语行为甚至比通常的行为主义更加不可取,因为不仅缺少运动行为,甚至没有活的身体外形,确切地说,心理属性不能归属于计算机(Dreyfus 1979)。另一种反对意见认为,即使计算机能够像图灵设想的那样去行事(诸如阅读十四行诗),它也并非真的具有智能,因为不能设想计算机真的会思考和理解:没有意向性,就没有智能。这一指责对于作为技术、甚至作为执行模仿的 AI 并不构成威胁,因为它承认完全具有人类特点的计算机性能是有可能存在的。此外,它并不否认,计算机模型在心理学中(像在其他科学中一样)可起到厘清理论的作用。但是它仍然坚持,AI 的概念内容不能帮助哲学家或心理学家去描述或解释心理过程本身,因为心灵具有意向性,而计算机没有,也不可能有。

在这种攻击中,一个颇有影响的例子是 J·塞尔的文章"心灵、大脑与程序"(第 3 章),该文运用图灵自己的计算概念来反驳图灵的观点:一台配有适当程序的计算机是有智能的。塞尔并没有把批评直接对准图灵的文章,而是指向该文章提出的两个纯理论条款:"强"AI(尝试通过编程构造真正的心理能力)和计算心理学(在这方面 AI 对心理学理论的内

容有所贡献)。

塞尔的第一个论点,包括他构想出来的“中文屋”,认为AI程序和计算机模型当然地是纯形式句法的(和图灵机一样)。基于这一点,他认为,一个系统不可能纯粹借助完成计算而达到理解。所以计算心理学决不可能解释我们的心理能力,任何一个程序更加不可能将智能赋予计算机。塞尔的第二个论点是,智能或意向性不仅需要心灵式的行为,同时还需要作为这一行为的基础的“正确的因果能力”。如前所述,对这样表达的论断,支持 AI 的人们是可以接受的。然而塞尔在定义这种因果能力时,根据的不是特性或功能,而是材料质别。此外,他还认为,从直觉上看,显然神经蛋白可以生成意向性,而金属和硅则不能。

我在反驳中指出(见第 4 章“逃出中文屋”): 即使最简单的程序也并不是纯形式主义的,而是具有某种相当本原的语义特性,所以从根本上说,计算理论并非不能解释意义。此外,只要大脑生成意向性的能力是清楚的,而不是完全反直觉的(头盖骨里面那些粘糊糊的物质如何可能进行理解?),这种认识所采用的信息加工方式同样可以用于计算机。这样,AI的概念就完全有理由被用作心理学理论的基本组成部分,同时某些想象之中的计算机也可以具有与意向性和智能十分近似的能力。任何一台计算机,即使它的内部计算组织与我们的完全等同,是否可以丝毫不含引喻地将其称为有智能的,则又是另一回事情,排除引喻不仅需要对事实的识别,还需要人类自身的道德评判(Boden 1987: 423 – 5)。

攻击图灵立场的第三条阵线(第 5 – 14 章与之有种种联系)持有的观点是: 与图灵的假定相反,要使计算机的表现在

深度、广度以及灵活性上与人类心智相媲美,在原理上和(或)实践上,都是不可能的。一台阅读十四行诗的计算机,不管是真有智能,或者仅仅只能模仿智能,都决无存在的可能。这种攻击所依据的常常就是图灵文章所驳斥的那些观点的变种:行为、创造性以及哥德尔定理的"非形式特性"(不能约简为规则的特性)的观点(Dreyfus 1979; Lucas 1961)。此外,技术性AI并非不能存在,实际上已经制造出实用的AI系统,但是AI和计算心理学的最高目标——人类心理过程的详尽的计算机模型——是不可能的,也是(或是)不可行的。

有些哲学家或许会反对说:这与不可行性无关,这里的问题是逻辑的可能性,而不是经验的可能性。这种回答忽略了"经验上可能的"在较为抽象的和较为实际的意义之间的差别,对于单独存在的基本科学原理和**受到非常普遍的现实世界的制约**的科学原理,应当区别对待。

图灵机在两种意义上都表现为经验上的不可能性,因为它有一个无限长的纸带。其他一些计算机器,就像传说中的地狱雪球一样,虽然没有被基本的科学原理所否认,但在现实世界中,由于时间和(或)空间的限制,也是无法实现的。例如,将野蛮搜索的算法用于下棋,每走一步都可能需要天文数字般的时间长度(虽然是有限的)。同样,即使加工单元的反应能够比神经元快得多,许多视觉任务也只有采用大规模并行处理方式,才可能在实际时间内完成。原则上讲,这种处理方式可以用一台串行计算机来模拟(所以某些理论问题能够在不规定串行或并行实现方式的情况下被提出来)。但是在实践中,只有相对比较小的并行系统才能如此完成。既然我们的大脑不是用天文数字般的时间工作的,我们就不

应当将我们的心灵归结为实施过程需要上千年时间的计算形式。现实世界进一步的限制是我们进化的起因，正如克拉克所说（Clark 1987a），如果某些类型的计算与进化是一致的，而另一些不一致，这就不仅牵涉到心理学，也牵涉到心灵哲学和 AI 哲学。

在那些与图灵一样相信 AI 可行性的人之中，无论在实践上使 AI 成为现实，还是在把它用于细化的心理学问题（既有抽象的任务分析，也有细致的实验观察）上，没有人比 A·纽厄尔和 H·西蒙做得更多了，正如"作为经验探索的计算机科学：符号和搜索"（第 5 章）一文所阐明的那样，在计算机同心灵哲学的关系上，没有人持更不妥协的态度了：心灵**是**一个计算系统，大脑**事实上是**在执行计算的职能（计算对智能来说是**充分的**），它与可能出现在计算机中的计算是**完全等同的**。人类智能可通过一组组控制着行为和（被逻辑上相似的行为主义心理学家所忽视的）内部信息处理的输入输出规则得到解释。由于计算机具备了正确的因果能力，它们也可以成为智能的：一台计算机——像一个大脑一样——是一个物理符号系统，而"一个物理符号系统具有对于一般智能行为来说，是必要的和充分的手段"。

纽厄尔和西蒙方法的核心，正如他们的反对者塞尔（第 3 章）所指出的，是与语义学的因果说相联系的符号论的形式句法理论。照他们看来，一个符号或计算成立的判据，是纯形式的，它的意义要借助于其因果的历史和作用来建立。一个符号就是一个物理**模式**，并以物理方式通过各种途径（如并列）同另一些模式发生联系，以构成复合"表达式"。（由物理手段实现的）计算过程可对**模式**进行比较和修正：一个表达式作

为输入，另一个作为输出。任何能够以物理方式存储并系统地变换表达模式的基底，都能行使符号的作用，但是这个基底与心理学目的无关。为达到对智能的理解，我们必须借助于指称和解释在信息处理层次上对物理符号系统加以描述。这两个语义学概念是从因果角度定义的，一个符号的意义就是这符号使系统产生的一组变化，或者达到或者响应某种（内部或外部）状态。因果相关本质上是任意的，就是说任何（非复合的）符号完全可以指称任何事物。〔这种限制不包括类比表述，因为在类比表述中，表述和被表述事物之间存在着不容忽视的相似性（Boden 1988：29－44）。〕

符号系统的这一定义可被批评为过分地物理主义，甚至那些在 AI 可行性方面与纽厄尔和西蒙有着共同信念的人，也有这种看法（Sloman 1986，待出版）。西蒙等人除了提出物质的例示对符号来说是**必不可少的**这样一个未加论证的假定（因此就排除了纯思维式的智能），他们的定义谈及的只是实际的机器，而不是虚拟的机器。所谓虚拟机，是指可被编程者看作正在使用的机器。它被抽象地定义为由有关系统执行的一组基本的信息加工作业。在虚拟机中，符号是抽象实体，而不是物理实体。一个计算系统之中可能存在若干个虚拟机——就像在高级编程语言由较低级语言实现的过程中，它依次被编译为汇编语言，然后再转换成机器码。心灵可能是由许多台这样的抽象符号机组成的，其中只有最基础的部分可通过脑组织以物质形式例示说明（这与某种层次较低的系统中的实际情况正相反）。然而，纽厄尔-西蒙的提法可以进行修改，以适应这一批评，批评者的观点并非认为 AI 不能实现，而是认为这比起他们两位定义中提出的那种文字

形式要复杂得多。

AI可以帮助我们认识心灵——它是什么,它是怎样工作的——这一点是所有计算心理学家都认可的,他们之中有些人同意纽厄尔和西蒙的看法:AI**就是**理论心理学(Longuet-Higgins 1987)。然而,也有些人对AI提出尖锐批评:D·玛尔在"人工智能之我见"(第6章)中指责许多AI工作与科学无关。AI程序常常建立在一些互不联系的、理论上无依据的见解和(或)毫无定则的经验摸索之上。在玛尔看来,智能科学需要建立在对基本(公理的)任务域的理论理解之上的"1型"模型,或"2型"的智能性能实现方式,产生后者的是"众多过程的同步行动,**这些过程的相互作用本身就是最简洁的描述**。"2型理论也许可以发现,也许(更可能的是)不能。如果一个任务所要求的2型解释太复杂,难以找到,这个任务就决不可能得到详尽的理解。只有在证明了1型理论不可能的时候,AI和心理学研究才会去寻求2型解释。

1型解释包括三个层次:"计算"理论提供规定有关信息处理任务内容的公理;"算法层次"描述能够执行该任务的步骤;"硬件"层次则指出算法是怎样实现的。玛尔像数学家一样,在使用"计算的"一词时,并不特指时间过程,而是指非时间的限制,时间过程是在算法层次上考虑的[前面有关虚拟机的讨论,已表明这不是一个单一层次,玛尔本人也假定视觉加工有若干层次(Marr 1982)]。如同每一任务都存在多种算法一样,每一算法也可能有不同的实现方式。对算法层次和硬件层次之间的相互制约所作的思考,有时可以使神经生理学家和计算心理学家共同从中获益(见第14章)。

(规定智能基本任务的)计算层次确立了AI和心理学的

自然属性——玛尔认为这一点在很大程度上尚属未知。他承认广义的2维至3维映射(低级视觉)和语法分析表现为自然属性,但是他把习俗心理学规定的种种任务排除在外(与社会心理学家研究颇多的意向属性一样),甚至也排除了被纽厄尔和西蒙模型化的算术技巧。在他看来,由于纽厄尔和西蒙忽略了基础的、无意识的信息处理任务,他们的工作缺乏科学性(对算术来说模式匹配可能是关键性问题,而他们却视模式匹配为理所当然的)。像其他的模块理论家(Fodor 1983)一样,玛尔否认"较高层次的心理过程"是可以详尽解释的,因为即使在医学诊断中,一定程度的自由选择也显然是存在的,何况在欣赏十四行诗时的浩渺无际的联想之中。因而,对专门知识或故事的理解作出科学的理解,或是为其建立有理解力的计算机模型,都是不可能的。技术性AI可能会产生有用的"专家系统",或许(某一天)也会产生出图灵的十四行诗阅读器,但对它的解释只能是一种过分复杂的、无法理解的2型理论。

一些计算主义者认为,玛尔关于科学解释的提法过于苛刻了。它忽视了对一套结构可行性进行阐释所起的解释作用,在这些可行性之中,自然现象必然存在,并且可以通过这些可行性系统地将这些现象加以比较(Sloman 1978)。而且,即使在1型分析有效的情况下(如低级视觉或语法分析),它同生物系统甚至人工系统的相关性也是有限的。进化过程在信手涂鸦中生成智能时,采用的不是设计角度,而是用现成材料"修修补补"。在有机体中可以看到许多精巧的工程作品,但并不奇怪,在计算机程序中也有与之不相上下的"拼拼凑凑"的非系统的方法(Clark 1987b)。实时限制表明(也

得到实验证据的肯定），即使视觉和语法分析有时也得益于权宜的处理方法，这种方法不能从1型方式中得到系统阐释，但是在缺乏详尽2型解释的情况下，它所起的作用是显而易见的。

有一种智能类型，关于它的1型计算的任务分析是否可能，引起了热烈的争论，这就是常识推理。AI中有一个使用广泛的假定：作为我们常识基础的思维是可以形式化的，或许甚至是可以演绎的。但是批评意见普遍认为，即使逻辑和（某些）科学推理能够通过规则模型化，日常思维却不行。以我们关于物理世界的默认的知识为例，它们是通过感觉运动的学习而获得的，与抽象的物理学原理毫无关系。根据这种观点，既然这种非语言化的知识不仅渗透到我们的运动行为中，而且渗透到我们的语言使用中，所以图灵设想的那种范围广阔的计算机式的对话是不可能的。

第7章介绍了这一争论的概况（对AI研究者所谓“框架问题”的介绍），第8、9章则对此作了详细讨论。对框架问题的认识源自机器人制定计划的语境，此外，该问题也出现在对语言理解和有关社会文化问题的常识性思维的AI研究中。该问题关系到对现实（物理的或社会的）世界中的行为后果的预见，这种行为后果可能是有意的，也可能是无意的、潜在起作用的。行为的形式表述必须明确地包括所有的有意结果，否则的话，有些结果就无法出现。只有当行为者很有运气，或者非常全面地对潜在起作用的外部情况作出明确预见时，行为的大量无意后果才会完全不起作用。框架问题能否得到解决，原则上取决于是否能以可信的方式对各种各样的行为后果作出足够彻底的明确表述。这个问题如何可能在实践中得

到解决(如果能解决的话),取决于相关的世界知识怎样才能被确认、表述和使用。

在"认知之轮:人工智能框架问题"(第7章)中,D·丹尼特简述了某些用于框架问题的AI方法,他指出,框架问题除了与AI的可行性有联系之外,还具有独立的认识论方面的重要性。哲学家们不承认它是一个问题,因为他们从未认真地提出这样的问题:他们偏爱的认识论原素**怎样**才能用于构建知识,以及根据知识进行推理。总之,很可能因AI研究的深入探讨而引申出新的哲学问题,以及对老问题的新见解。因为AI研究者在编写程序时无法忽略(尽管他们有时会回避)"怎样"的问题。即使他们像某些哲学家那样,不是以编写程序为目的,而是要为已知的任务域提供(1型)内容分析,他们对未来任何一个程序运转所不可缺少的明确性的重视,也会使他们的抽象分析从中获益。

很多有关框架问题的研究都属于后一种类型,带有抽象的特点,与哲学中的逻辑和认识论十分相近,但是它同心理学的亲缘关系却存在两方面的疑问:第一,它不必刻意模仿心理的真实状况,它的目标不过是为机器人提供一个可使用的、可靠的常识性知识表述方式,与人的头脑里进行的是何种**过程**无关;第二,常识性知识的**内容**是否能够公理化,甚至形式化,关于这个问题,在AI内部也存在着某种分歧。

在常识的形式化特性问题上,P·海斯与图灵有着相同的信念,他(与J·麦卡锡一起)首先确认了框架问题。在"朴素物理学宣言"(第8章)中,他(像AI批评家们所评论的那样)认为,对物质世界所作的日常思维,并不需要采用理论物理学——至少在解决相关的信息加工任务时,理论物理学是不适

用的。相反,这种思维采用的是“朴素物理学”——我们关于环境的那些未经教授而得的、基本上是无意识的知识。在感觉运动技能(如倒水)和(对动词“倒水”的)语言理解中,就包含着这种知识。朴素物理学的形式化需要对涉及物质、原因、空间和时间的许多概念的分析,海斯〔在这里及后继文章中(Hayes 1985)〕提出的概念还有:**重量**、**支撑**、**速度**、**高度**、**内部**、**外部**、**相邻**、**边界**、**路径**、**进入**、**障碍**、**流体**和**原因**。同样,海斯认为我们在实践上和语言上对社会生活的理解依赖于“朴素物理学”,它不是由从经验上归纳人们如何行动构成的,而是由决定着日常心理能力的基础概念和推理模式构成的。AI如果想要解释智能,或是得到阅读十四行诗的计算机程序,必须首先完成朴素物理学分析和朴素心理学分析。

这一艰巨任务可以用不同的方式来完成。一是“逻辑主义”的方式,海斯的研究就是最好的例子,其观点是:我们的基本的常识性知识不仅可以形式化(用某种形式语言,如编程语言,表示为形式得当的公式),而且可以公理化。公理表示的是物理世界和社会现实中的普遍真理,从它们出发,经过正确的推导,就会得出像逻辑定理一样可靠的结论。的确,它们也有可能用形式逻辑来表示(例如谓词演算)。逻辑主义AI的研究目标,首先是形成一个有关常识基础的抽象(1型)理论,其次才是编写程序,把这种知识用于机器人学,制定计划或是作语言理解。

多年来,海斯可以把D·麦克德莫特看作是逻辑主义者中的一员,但是最近,麦克德莫特公开放弃了原有主张,在“纯粹理性批判”一文(第9章)中表示了自己的醒悟。麦克德莫特的论述使人想起与知识可形式化特性不同的观点(第13章),

但是他并未改变对可形式化特性的看法,他怀疑的是对知识公理化,而不是对知识编程。

麦克德莫特认为,逻辑主义的致命错误是把演绎和计算混为一谈,因而假定所有思维本质上都是演绎的。包括他本人在内的逻辑主义者,为定义"非单调"逻辑学——这种方法可以通过增加新信息(作为附加前提)而取消演绎结论——所作的努力,不能处理日常偶发事件。因为公理化(1型)分析仅限于演绎范围,所以非演绎的AI程序只有借助于非演绎推理的一般理论,才能科学地加以认识,而非演绎推理正是哲学家们寻找已久而未能发现的东西。麦克德莫特承认存在这种可能性:AI的许多研究可能都是没有出路的,"只是因为对哲学家们昔日的失败一无所知,才得以维持"。但是他也希望,AI的概念和技术有助于我们发现传统认识论所没有发现的一般性理论。在此之前,不会出现无所不包的智能科学,我们为非演绎AI程序所做的辩护只可能是:"它有效!"

常识性思维可能给AI带来困难,但它至少是思维,是一种大多数人都认为,在所有事物中最适于建立计算机模型的心理活动。动机和情感则是另一类事物,人们普遍怀疑能够通过计算途径来模拟或解释心灵的这些方面(Dreyfus 1979; Haugeland 1978)。这种怀疑不仅与它们的意识维度有关——因为推理也可以看作是有意识的——而且也与它们的一般本质特征有关。动机(以及其他意动范畴,如意向)是行为的源泉或驱动力,同人格与自我有十分密切的关系,它们怎么能同计算相提并论呢?拿情感来说,它们正好与理性相反,会导致我们以在较为冷静时难以接受的方式去做、去看一些事情。尽管认知能够通过计算方式来认识,情感就肯定不能吗?对

计算分析来说，情绪看起来更难对付，事实上，它对我们所有的思维、行动和经验都产生着影响（看来它与在整个身体中扩散的化学物质有关，或是由这些物质引起的）。总而言之，意向和感情可以通过计算方式而变成理论上可理解的，这种看法似乎有点不可思议。

对这种认为非认知现象是不能通过 AI 方式来认识的信念，一些计算主义者抱有完全不同的看法，他们认为：任何智能系统，如果有一些相互独立、又潜在冲突的目标，并且在复杂快变的环境中行动，那么它就需要一些为动机和情感所固有的内部控制机能。所以只要 AI 研究者们打算建立智能模型（或充分理解智能），他们就必须也建立动机和情感模型（认识动机和情感）。这种观点是从设计角度着眼的，因此问题就变成：如果要使某种计算系统成为可能，那么（工程师或是进化过程）必须提供什么样的特征。例如，我在别处论证过，人本心理学家采用的许多理论概念（包括动机、意向、情感、情绪、性格和个人理想）明显表现出控制和组织特征，它们对于多目标系统来说是不可少的（Boden 1972，1973）。设计角度的思想方式同样深深地影响着 A·斯洛曼关于“动机、机制和情感”的讨论（第 10 章）。

斯洛曼在概述“可能心灵空间”的某些维度时指出，多目标系统从本质上说会导致内部冲突，其解决方式需要特殊类型的控制机制，同时又指出，由于情感与某些这样的机制相关联，所以它并不是心灵的一个特殊的子系统，而是一种无处不在的心灵特征。一个具有多种动机，但时间资源、精力资源和知识资源都有限的自主系统，需要有一些在动机之间进行比较和选择的策略，从而决定要做什么。由于这种决策常常影

响到进一步的动机(意向)的产生,所以这种系统的内在特征是递归的。爱好和道德评价代表着进行比较和选择的背景判据,它们能节约计算劳力,因为它们不必在每一决策时刻都重新生成。关乎生存的动机可能需要打断正在进行的活动,方法是即刻强行占用有效资源(工作记忆、肌肉),在未经任何广泛比较或没有任何决策过程的情况下,自发地采取行动。不同的情感与各种类型的动机控制有关:与恐惧相关的是因觉察到危险而中止行动,与焦虑相关的是追求重要目标时觉察到有成功的可能,等等。

动机、情感和意向是与大脑相关联并且最终在大脑中完成的多层虚拟机器集合的一些方面。斯洛曼指出,他的论述所谈的是(有多重目标而资源有限的)相关类型的计算系统,至于它们是由冯·诺伊曼机还是由并行机来实现,是无所谓的。但是,尽管冯·诺伊曼计算机(图灵机的近似形式)在原则上可以执行任何计算,要在实际时间里起作用,某些计算很可能需要由类型根本不同的方式来实现。毫无疑问,实现动物心理过程的机器(大脑)与计算机有着很大的差别。

联结论就是受到这一事实启发而形成的 AI 分支(第 11 – 13章)。联结论是一个总的名称,其中包括多种信息处理系统,它们不同的计算特性还远未得到认识(Anderson and Rosenfeld 1988)。它们的共同特征是,通过概念化,它们可以成为由许多简单单元构成的大规模并行处理装置。一个单元的活动受到相邻单元活动的制约,相邻单元通过抑制或激活连接方式与之发生联系,其连接强度可以根据设计和(或)学习而发生变化。单元活动和连接强度可用数字表示,整个系统活动和系统加权值的变化(通常)是由微分方程控制的。

这些单元可以是各种不同类型的：二元的（活动或静止）或连续梯度的（活动程度是变化的）；确定的（其活动完全取决于输入单元的活动）或随机的（有时随意激活）；专用特长的（有效证据只对一个单元产生影响，对其他单元不起作用）或作用重叠的（对部分共享证据作出反应）；可能协调得比较一致，也可能差些。这些计算上的差别，全都可以在大脑中找到例证。然而，尽管从神经系统中得到很多启发，联结论单元——像麦卡洛克和皮茨在理论上所规定的“神经元”——仍然是抽象的理想形式：实际的神经元不仅复杂得多，而且有某些明显不同的特性。

联结论的一个很有影响的例子，是对并行分布式处理所做的“PDP”[①]研究，其中包括G·欣顿、J·麦克莱兰和D·鲁梅哈特所描述的“分布式表述”（第11章）。在PDP系统中，一个概念不是由存储在某个可确认的存储部位中的个别符号表述的，而是由一个均衡状态表述的，该状态定义在由局部相互作用的单元构成的动态网络上。每个单元对与所涉及概念相关的许多微观特征之一进行编码，单元之间的联结则根据响应特征是相互支持的还是相互抵制的，而表现为激活的或抑制的。在均衡状态下，那些高度活动的单元代表相互支持或至少是相互一致的特征（其他单元则处于静止状态）。任一给定的单元都可以参与表述若干个概念，在不同背景下，“完全相同的概念”可以用局部不同的网络来表述。

分布式表述的内在特性，使它能够完成某些很难用传统方式编程的计算。模式匹配即为一例，即使新模式与旧模式

① PDP即Parallel distributed processing（并行分布式处理）的缩写。——译者

有某些差异，或者只是旧模式的一部分，匹配也能完成。因此，知觉识别、类比思维和族的相似性分类——对传统 AI 来说，每一个都是重大的障碍——对 PDP 计算来说都不成为问题。PDP 系统自然而然地具备了这些能力，不必通过事先规定大量个别规则，专门编程输入这种能力（虽然还得具有表示潜在基本特征的单元）。哲学家曾经引用这些心理能力，以论证不可能规定一组囊括所有这些能力的规则，因而 AI 不可能对它们作出解释。这种说法其实是把“AI”限定在它最早的形式之中，那时，每个程序符号表述的是某个可确认的概念，而不是“亚符号式”的微观特征。

关于传统 AI 和联结论 AI 之间的理论关系，存在着争议，即使是在联结论者内部也是如此。当然，所有联结论系统原则上都可以由冯·诺伊曼机来实现。但是若撇开两种 AI 形式的这种非常抽象的原则上的等价不谈，关于它们在认知科学中的相互关系和用途对比，在看法上是有分歧的。正如 A·克拉克在“联结论、语言能力和解释方式”（第 12 章）一文中所指出的，这些争议涉及两点：每一种方法建立特殊心理能力模型的适宜性，以及联结论性能模型的解释性地位。

就第一点而言，忠实于传统论的人往往认为，语言理解和分步推理所要求的计算形式，并不完全适合于联结论系统。作为联结论者，第 11 章的作者也同意这一点，他们认为，为了完成这些任务，PDP 系统也许不得不模拟冯·诺伊曼机。也就是说，那些能有效完成这些计算的虚拟机是属于传统 AI 描述得最好的那种，虽然在大脑中这些计算的基本实现方式是联结论式的（参阅 Smolensky 1988; Clark 1989）。从这个意义上说，AI 的两个分支是互为补充的。

关于第二点,说法各异。例如,常有这种看法:联结论并不涉及心理过程,而是涉及心理过程的神经实现方式。然而,尽管联结论在总的方面是关心生物学上的可实现性的,但是大多数联结论系统所作的并不是神经实现方式的模型,而是抽象定义的信息处理的模型。只有表现特殊神经回路和(或)突触相互作用的计算系统(如第 14 章中),才是这个意义上的模型实现方式。

强调这第二点的另一方式是询问联结论模型提供的是何种类型的解释。它们能够体现被玛尔(当作"1 型"理论)以及纽厄尔和西蒙(当作"知识层次"理论)提出的那些作为不同心理领域基础的抽象原理吗?或者它们不过是一些在很大程度上难以理解的存在证据,是一些这样的模型,它们通过参照体现在它们之中的均衡和学习的抽象原理(如波尔兹曼方程或反向传播)而获得支持,其功能充其量也只能接受一种 2 型解释(列出所有联结的加权值)吗?这样的模型或许会有力地增强技术性 AI,并且,可以想见,它会有助于产生图灵所设想的智能型计算机性能。但是,它们能为智能科学本身提供有用的解释吗?

克拉克认为:联结论并不局限于 2 型解释,但也不是建立在 1 型解释的基础上。联结论系统不能作为例证说明由纽厄尔和西蒙以及由玛尔所建议的、表现为基本公理的、解释性的"上下贯通形式(cascade)"。当然,一个抽象的任务分析,可以用来规定相关的、作为基础的输入和输出,甚至用来(在非学习模型中)配置许多联结加权值。例如,一个作出低级视觉模型的系统,可能具有响应特殊类型 2 维信息的输入单元,以及计算特殊类型 3 维信息的输出单元,这种选择完全取决于

2维至3维的映射理论。对联结论系统所做的事情来说，语言能力理论甚至可能是一种有用的理想形式，但是它不能解释实际上正在进行的是什么，因为联结论系统既不含有1型理论的显式知识，甚至也不含有1型理论的默认知识。总的来说，在联结论系统所作的信息加工和完成“相同”任务的冯·诺伊曼系统之间，不存在精确的对映。

克拉克指出，联结论涉及传统认知科学的“方法论转换”。在着手作出存在于已知智能形式中的加工过程的模型之前，联结论并不是，也不需要先寻求公理化的任务分析。相反，他们建立的网络只需借助于抽象语言能力作出松散的说明（在“0.5级”上），他们让该网络学习完成当前的任务，然后他们才发现该网络已开始体现的这些高层原理。模型的解释力度取决于这些原理，而不取决于那组（很可能是难以理解的）联结加权方式。这些高层原理究竟是什么样的，尚不清楚。为了揭示它们，可采用的方法有对网络症状的研究（探求故意对系统造成的变化会产生什么影响），激活方式的记录（从而确认在特定时间或相继时间里活动的单元），以及簇分析（它揭示出已知系统中激活模式的层级结构）。例如，作为现实世界活动基础的概念结构，或许可以用这些方法按照**发生于其后者必然是其结果**（post hoc）的规划来揭示。这与朴素物理学中的逻辑主义策略——编制一个预设的、主要用于**此时此地某一特定目的**（ad hoc）的公理表的做法恰恰相反。

H·德雷福斯和S·德雷福斯在“造就心灵还是建立大脑模型：人工智能的分歧点”一文（第13章）中承认，对于那些耳熟能详的批评，联结论可以在一定程度上为AI作出辩护。但是他们仍然反对海斯和克拉克的观点，坚持认为，语言和常识

不可能为 AI 所获取(即使是联结论类的也不行)。在捍卫这一立场时,德雷福斯把 AI 研究同内容广泛的哲学文献联系起来,比较了西方理性主义传统同大陆的现象学和后期维特根斯坦的不同特点。他对 AI 的怀疑源于这样的观点:人们并不使用关于日常世界的理论,科学也不能表达这样的理论,因为并不存在一组与语境无关的理解原素。我们的知识是熟练的**技能**,不同于过程规则、表述方式或**知识内容**,甚至我们关于形式系统的知识也不能不借助于有关怎样延续数学序列或应用逻辑规则的背景直觉。

德雷福斯在概述 AI 两个分支激烈对抗的历史(但略去了作为这两个分支共同来源的麦卡洛克和皮茨的工作)时谈到神经网络研究的早期成就和暂时沉寂,形式主义 AI 的初始成功,以及继之而来的、对常识性理解编程时遇到的困难。特别是,传统 AI 不能获得整体知觉、语境敏感性和对族的相似性及相关性的识别——而这每一方面在联结论中都得到了较好的处理。德雷福斯对 AI 前 20 年评价是"理性主义传统终于被置于经验检验之下,而它失败了"。德雷福斯认为联结论是对他这一观点的支持:智能并不依赖于关于世界的理论,所以不能在计算系统中通过规则来获取,或作出模型(Dreyfus 1979)。

我们或可反对说,联结论研究的也是计算系统,其单元是根据精确规定的过程和规则进行计算的。德雷福斯的回答是:由联结论单元计算出的函数,一般是相当抽象的,以致它们无法直接与可用语言表达的概念或信念发生联系,而且它们甚至常常得不到有关科学家的确认(参阅第 12 章)。此外,一个单元的活动性和影响并不是由单个指令或规则决定的,

而是随其他单元的活动性变化的，这就使得任何试图把联结论单元看作是带有确切语义含义的理论原素的做法，都失却了稳固的基础。同时，联结论“规则”（微分方程）也不同于传统 AI 所钟爱的那种逻辑主义公理。

但是德雷福斯对智能理论仍然感到失望——即使是建立在联结论信息加工基础之上的理论。他认为，只有大小和回路结构与大脑近乎完全等同的联结论系统才能作出人类智能的模型，这种系统还要具备人类动机、文化追求和血肉躯体。

我们主张，比起 AI 迄今为止的做法，它应当更仔细地关注大脑，但这并不等于说我们赞同上述观点（AI 应当做出整个大脑的模型）。尽管当代联结论发轫于关于“神经活动”的概念，但它却很少借鉴神经科学。（传统 AI 借鉴得更少，它认为神经实现方式的问题与计算问题不仅有区别，而且全然无关。）在有关低级视觉的 AI 著作中，有时提出或借用关于视神经解剖学的假设。但是一般说来，当代联结论的目标是确认抽象定义的（理想化的）联想网络的计算特性，而不是作出实际神经元复杂性的模型，或对映实际的神经回路。实际上，我们甚至还不清楚，实际神经元的哪一部分或哪几部分，与联结论模型的单元和联结有着最接近的对应关系（Smolensky 1988：§4）。

神经科学家特别关心实际的神经元和神经解剖学，他们常常利用建立计算模型来揭示，一个已知的（单细胞的或多细胞的）神经解剖学结构，怎样使得大脑能够加工某些特定类型的信息。在神经科学家迄今已经设计出的神经元计算机模型中，有许多都是特别注意未加解释的细胞激活方式，而不是用心理学方式解释的神经活动。然而，尽管处在具有跨学科特

点的认知科学中心位置上的神经科学家(还)为数不多,而自认为是在作AI研究的人就更少了,但是他们所作的研究,是总的智能科学中最基础的部分,同时这种研究与AI思想的共同之处也在不断增加(Thorpe and Imbert,待出版; P.S.Churchland 1986; P.M.Churchland,待出版)。例如,玛尔不仅在作出视觉模型时,而且在作出运动控制模型时,都借鉴了神经科学(他的小脑理论采用了某些近期联结论模型中所用的学习规则的一个较早的版本)。

神经解剖学的详尽知识可能会提出一些计算系统类型,它们不同于传统AI或联结论AI中所研究的系统。P·丘奇兰在“认知神经生物学中的某些简化策略”一文(第14章)中,概述了这样一个例子。它与冯·诺伊曼式的或联结论的构造体系不同,含有按照特定大脑结构作出模型的一些相互联系的神经元层,它的作用不是符号加工或达到平衡,而是进行坐标变换。“状态空间分层结构”这一基本思想出现在论述各种感觉“2维至2维”变换的神经科学著作中。能够定义多维状态空间的、更抽象的“神经矩阵”也已形成,它可用于解释小脑是怎样计算运动协调的(“躯体经验”的一个方面)。

这些思想经丘奇兰精心设计,被推广到其他心理学领域,其中包括对颜色和味道的细微辨别。这里,“辨别”也可以用“经验”来代替。丘奇兰的哲学目标是,证明消除性的唯物主义是可信的;其做法是描述一种神经计算方法,由它说明这一点如何可能实现。(正如他所指出的,有理由认为,这个方法更加适合于计算孤立的感觉辨别和感觉运动协调,而非结构化的语言理解或随意动作。)据消除性的唯物主义的看法,未来的神经科学或许能提供有关我们知觉和思维的信息,从而

使我们不再把心理事件(甚至感觉特性)看成与大脑状态有什么不同。为了做到这一点,神经科学家首先必须详细地说明,我们的心理状态怎样才能成为可能。值得注意的是(请塞尔原谅),关于怎样才能做到这一点,丘奇兰所提的建议,不是借助神经生物化学得出的,而是根据大脑回路得出的,这种回路能够实现特殊类型的信息加工。从 AI 的观点来看,这正是我们所期待的。

由此可知,这十四篇文章提供了范围广阔的 AI 哲学述评。它们表明,AI 拥有各种迥然不同的方法论,就像数学囊括了各种不同类型的理论一样。由于最早的 AI 分支(GO-FAI)中存在着缺陷,就把 AI 贬斥为哲学赝品而置之不理,这好像 17 世纪的哲学家,因为伽利略没有微分方程而无法解释流体动力学的特性,就摈弃他关于"数学是上帝的语言"的见解一样。虽然 4 个世纪以来的物理学已证明了伽利略的正确性,但是可以肯定,把 AI 说成是一般性的智能科学,到目前为止只不过是提供了一张期票而已。在评价这一承诺时,我们不仅要相当仔细地考察当前的 AI,而且应当记住,科学一般说来不是"纯经验的"。科学可以提供新的观念,使哲学问题新的提问方式成为可能——哪些问题可能在这一过程中发生转化,因为科学进步会引起深层概念更迭(Putnam 1962; Churchland 1979)。支持 AI 的人们认为,通过这种方式,AI 就能够对心灵哲学和认识论作出贡献。正如丹尼特(Dennett 1988)所指出的:"AI 尚未揭开任何古老的心灵之谜,但是它为我们提供了规范和拓宽哲学想象力的新方法,至于对这些方法的利用,我们还刚刚开始。"

编者附言:第 15 章是一篇以前未曾发表过的文章,作者

A·屈森斯。它是在本书已经付印时编入的。

屈森斯认为,GOFAI 不能解释意向性,也不能说明心灵在物理世界中的显现。形式主义 AI,以及提出"思维语言"的心理学理论,都涉及到概念之间的(句法)关系,而概念的存在,以及语义特性,又被看作是理所当然的(Fodor 1975, 1981:"方法论的唯我论"章)。

为了解释概念的显现,我们需要一个非概念论的心理内容的观念,一个关于这种内容怎样才能出现在(非形式主义的)表述系统之中的说明,和一个关于概念怎样才能逐步由它构造而成的分析。为了解释心灵/世界的区别,或是客观性,我们必须说明,概念表述怎样才能具有像"参照"、"真"和"假"这样的语义特性。

屈森斯根据最近的哲学语义学和形而上学著作,给基于经验的非概念性内容观念下了一个定义("构造理论内容",即CTC),并分析了物理系统具备一个概念是怎么回事。他认为,概念原则上可以由 CTC 构造,同时保真推理有赖于概念的内部结构,而不是概念之间的句法关系。至于概念会怎样在实践中出现,联结论者研究了视角依赖逐步降低的表述方式的构造〔Marr 1982,也可见 Hinton"并行系统中的形状表述", Proc. Seventh IJCAI(1981), 1088 - 96〕,指出一个心理系统怎样能逐步形成客观性。

参考书目

Anderson, J. A., and Rosenfeld, E. (1988). *Neurocomputing: A Reader*. Cambridge, Mass.: MIT Press/Bradford Books.

Boden, M. A. (1972). *Purposive Explanation in Psychology*. Cambridge, Mass.: Harvard University Press.

—— (1973). 'The Structure of Intentions.' *J. Theory of Social Behaviour* 3: 23–46.

—— (1987). *Artificial Intelligence and Natural Man*, 2nd edn., London: MIT Press; New York: Basic Books.

—— (1988). *Computer Models of Mind: Computational Approaches in Theoretical Psychology*. Cambridge: Cambridge University Press.

Churchland, P. M. (1979). *Scientific Realism and the Plasticity of Mind*. Cambridge: Cambridge University Press.

—— (in press). *The Neurocomputational Perspective*. Cambridge, Mass.: MIT Press/Bradford Books.

Churchland, P. S. (1986). *Neurophilosophy: Toward a Unified Science of the Mind–Brain*. Cambridge, Mass.: MIT Press/Bradford Books.

Clark, A. J. (1987*a*). 'Connectionism and Cognitive Science.' In J. Hallam and C. Mellish (eds.), *Advances in Artificial Intelligence*, pp. 3–15. Chichester: Wiley.

—— (1987*b*). 'The Kludge in the Machine.' *Mind and Language* 2: 277–300.

—— (1989). *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing*. Cambridge, Mass.: MIT Press/Bradford Books.

Dennett, D. C. (1988). 'When Philosophers Encounter Artificial Intelligence.' *Daedalus* (Winter 1988): 283-95. Also published in S. R. Graubard (ed.), *The Artificial Intelligence Debate: False Starts, Real Foundations*. Cambridge, Mass.: MIT Press, 1988.

Dreyfus, H. L. (1979). *What Computers Can't Do: The Limits of Artificial Intelligence*, rev. edn. New York: Harper & Row.

Fodor, J. A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, Mass.: MIT Press/Bradford Books.

Haugeland, J. (1978). 'The Nature and Plausibility of Cognitivism' (with peer-commentary and author's reply). *Behavioral and Brain Sciences* 1: 215–60.

—— (1985). *Artificial Intelligence: The Very Idea*. Cambridge, Mass.: MIT Press/Bradford Books.

Hayes, P. J. (1985). 'The Second Naïve Physics Manifesto.' In J. C. Hobbs and R. C. Moore (eds.), *Formal Theories of the Commonsense World*, pp. 1–36. Norwood, NJ: Ablex. Repr. in R. J. Brachman and H. J. Levesque (eds.), *Readings in Knowledge Representation*, pp. 467–86. Los Altos, Calif.: Morgan Kaufmann.

Longuet-Higgins, H. C. (1987). *Mental Processes: Studies in Cognitive Science*. Cambridge, Mass.: MIT Press/Bradford Books.

Lucas, J. R. (1961). 'Minds, Machines, and Godel.' *Philosophy* 36: 112–27.

McCulloch, W. S. (1965). *Embodiments of Mind*. Cambridge, Mass.: MIT Press.

Marr, D. C. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: Freeman.

Minsky, M. L., and Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. Cambridge, Mass.: MIT Press.
Putnam, H. (1962). 'Dreaming and Depth Grammar.' In R. J. Butler's (ed.), *Analytical Philosophy*, pp. 211–35. Oxford: Blackwell.
Sloman, A. (1978). *The Computer Revolution in Philosophy: Philosophy, Science, and Models of Mind*. Brighton: Harvester Press.
—— (1986). 'What Sorts of Machines Can Understand the Symbols They Use?' *Proc. Aristotelian Soc.*, Supp. 60: 61–80.
Smolensky, P. (1988). 'On the Proper Treatment of Connectionism' (with peer-commentary and author's reply). *Behavioral and Brain Sciences* 11: 1–74.
Thorpe, S., and Imbert, M. (in press). 'Neuroscientific Constraints on Connectionist Modelling.' In R. Pfeiffer, Z. Schreber, F. Fogelman, and T. Bernold (eds.), *Connectionism in Perspective*. Amsterdam: North-Holland.
Turing, A. M. (1936). 'On Computable Numbers, with an Application to the *Entscheidungsproblem*.' *Proc. London Math. Soc.* 42: 230–65; also 43: 544.

1 神经活动内在概念的逻辑演算

W·S·麦卡洛克和 W·H·皮茨*

1. 引　言

理论神经生理学建立在某些基本假定上。神经系统是一个神经元网，每个神经元都有一个细胞体和一个轴突。它们的附属部分，或称突触，总是位于一个神经元的轴突和另一个神经元的细胞体之间。神经元任何时刻都有某个阈值，刺激必须超过这个值才能激发起一个冲动。这一过程，除了它出现的这个事实和出现的时间外，都是由神经元而不是由刺激决定的。这个冲动从刺激点传播到神经元的所有部分。沿轴突传播的速度与轴突的直径直接相关，变化范围从在细的（一般也是短的）轴突中小于每秒 1 米，到在粗的（一般也是长的）轴突中大于每秒 150 米。因此，在确定同一来源的冲动到达不同远程点的时间时，用于轴突传导的时间就显得不怎么重要了。占主导地位的是刺激通过突触从一些轴突末梢到达一些细胞体。这一点究竟是取决于个体突触的不可逆性，还是仅仅取决于大量存在的解剖学构形，仍是一个悬而未决的

问题。若假定后者,则无需特别的假设,并且能解释许多已知的例外情况,但是任何关于原因的假定都要与进一步的演算相容。已知这种情况是不存在的:在任何一个神经元中,通过单个突触的刺激已经诱发出一个神经冲动,而这个神经元又可能在小于四分之一毫秒的不易察觉的附加期里,再被数目足够多的相邻突触传来的冲动所刺激。在较长的时间段上观察到的冲动时间累加,在单个神经元里是不可能出现的,这实际上取决于神经网的结构特性。冲动到达一个神经元与神经元本身传播的冲动之间,存在着大于半毫秒的突触延迟。在神经冲动的开始阶段,神经元对于任何刺激作用都毫无反应。随后,它的可刺激性迅速恢复,在某些情况下达到正常值以上,再从这点降落到低于正常值,然后慢慢恢复到正常。频繁的活动增加了这种低于正常值的特性。这种神经冲动所拥有的特性,仅仅取决于神经冲动的时间和位置,而与任何别的神经能的特性无关。近来,经过严格论证,只有抑制是与这一理论相抵触的。抑制是神经元的一组活动被第二组并发的或先发生的活动所终止或阻止。这一点直到最近才能根据如下假定获得解释:第二组神经元的预先活动可能会提高中间神经元的阈值,使得它们不再能被第一组神经元所刺激,而第一组的冲动必须加上这些中间神经元的冲动,才能刺激正处于抑制状态的神经元。今天,我们已知某些抑制所耗费的时间

---

* W·S·麦卡洛克和 W·H·皮茨,"神经活动内在概念的逻辑演算",见 W·S·麦卡洛克,《心灵的实现》,麻省理工学院出版社,1965,第 19 – 39 页。麻省理工学院出版社允许重印。

W·S·麦卡洛克(Warren S. McCulloch),精神病学家和神经生理学家,麻省理工学院电子研究实验室成员。W·H·皮茨(Walter H. Pitts),麻省理工学院数学系讲师。

小于一毫秒。这就排除了中间神经元而需要突触，冲动通过突触抑制了那个正在被来自别的突触的冲动所刺激的神经元。现有的实验还未表明这一无反应性是相对的还是绝对的。我们假定是后者，并证明对我们的论点来说这个差别是无关紧要的。任何种类的无反应性都可由下列两种方式之一来解释。“抑制性突触”可能是那种产生提高神经元阈值物质的东西；或者它的位置可能使它的刺激所产生的局部扰动抵制另外一些刺激性突触所诱发的变化。鉴于我们已知在电刺激场合时位置也有这种作用，所以在尚未得到证实之前，我们将排除第一种假设，因为第二种假设不包含任何新的假设。这样，我们就在同一总前提下得到两种关于抑制的解释，不同之处仅在于所假定的神经网，以及由此造成的抑制所需的时间。以后我们把这种神经网作为**广义等价物**来看待。既然我们所说的网的特性在等价条件下是不变的，我们就可以采用演算起来最方便的物理假定。

很多年前，本文作者之一另辟蹊径，就这一论点得出这种设想：任一神经元的响应事实上都等价于提出了一个使神经元受到充分刺激的命题。于是他试图用命题的符号逻辑标记来记录复杂神经网的行为。神经活动的“全或无”规律足以确保任一神经元的活动可以表述为一个命题。神经活动中存在的生理关系当然是与命题中的关系相对应的；表述的功用取决于这些关系与逻辑命题关系的等同性。对任一神经元的每个反应，都存在一个对应的简单命题陈述。而这又意味着，根据当前神经元上的突触构形和该神经元的阈值，或者得出另外某个简单命题，或者得出类似命题的析取或合取，否定形式的，或非否定形式的。两个难点出现了。

第一点与助长和消退有关,在这两种变化中先前的活动暂时改变了该网同一部位对后继刺激作用的响应方式。第二点与学习有关,在学习中以前某个时刻并发的活动使网发生了永久性改变,以致于原先不充分的刺激现在充分了。但是对于经历这两种变化的网来说,我们可以代之以假想的等价网,而替代网是由连接形式和阈值都未经改变的神经元组成的。但是有一点必须澄清:我们谁也不认为这个形式上的等价物就是实际的解释。恰恰相反!我们认为助长和消退依赖于阈值的连续变化,它与电变量和化学变量有关,例如后电位和离子浓度;而把学习视为永久性变化,这种改变在入睡、失去知觉、痉挛和昏迷后仍能保持。形式等价物的重要性在于:作为助长、消退和学习的实际基础的变化,决不影响从对神经网活动的形式处理中得出的结论,并且对应命题的关系仍是逻辑命题的关系。

神经系统包含许多环形通路,它们的活动可再现任一参与神经元的刺激,因而使得关于经历时间的参照变得不确定了,不过,它仍表明传入神经活动已经实现了一个属于某种类别的时间上的构形。通过递归函数对这些内在关系作出准确说明,同时确定出那些能够在神经网活动中实现的内在关系,我们就完成了这一理论。

2. 理论:无环网

我们对演算作如下物理上的假定:

1. 神经元的活动是一个"全或无"的过程。

2. 为了在任何时刻刺激一个神经元,在潜伏的附加期内必须有一定数目的突触受到刺激,而这一数目与这个神经元以前的活动和位置无关。

3. 神经系统内唯一有效的延迟是突触延迟。

4. 任一抑制性突触的活动都完全阻止了那一时刻的神经元的刺激。

5. 网的结构不随时间变化。

为了描述这一理论,最适用的符号系统是 R·卡尔纳普(Carnap 1938)的语言 II 系统,并以 B·罗素和 A·N·怀特海(Russell and Whitehead 1927)的各种记号,包括有关点的基本惯例,作为补充。但由于印刷上的需要,我们只好使用正的而不是倒的"E"来表示存在算子,用箭头("→")而不是马蹄型符表示蕴涵。我们还将使用卡尔纳普的句法记号,但用黑体字型而不用德文字型印刷;我们将引入一个函子 S,对于特性 P,它的值表现出这种特性:当 P 获得 P 的先行值时,它就获得一个数;它由"$S(P)(t) . \equiv . P(Kx) . t = x$"定义;自变量的括号经常被省略,在这情况下,就被理解为右边最近的谓项表达式[Pr]。此外我们还把 $S(S(Pr))$写成 $S^2 Pr$,等等。

一个已知网 H①的各个神经元可以分别记为"c_1","c_2",…,"c_n"。继而,在神经元 c_i 从时间原点起经过一定的突触延迟数后在某一时间激发时,我们将用"N"来表示这个数的特性,并以数字 i 作为下标。因此 $N_i(t)$就表明 c_i 在 t 时被激发。N_i 称为 c_i 的动作。有时我们会把"N"的下标数看作

① 此处和以下以字母"H"替代原文中德文字母 N 的花体字型。——译者

像是属于对象语言一样，并代表一个函子的自变量，所以它可以用数值变量[z]替代并量化；这样我们就能用一个算子来缩写长的但是个数有限的析取和合取。在 Pr 序列中我们将十分普遍地采用这种习惯用法；可以用一个显而易见的析取定义从形式上使它得到保证。谓项"N_1"，"N_2"，…，构成了句法类别"N"。

我们把 H 的**周围传入神经**定义为 H 的神经元，而 H 不带有对其产生突触作用的轴突。设 N_1，…，N_p 表示这种神经元的动作，而 N_{p+1}，N_{p+2}，…，N_n 表示其余神经元的动作。于是 H 的**解答**将是一类 S_i 形式的语句：$N_{p+1}(z_1) .\equiv. Pr_i(N_1, N_2, \cdots, N_p, z_1)$，式中 Pr_i 不包含除 z_1 以外的自由变量，也不包含除自变量[Arg]中的 N 以外的描述符号，而可能包含某种常量语句[sa]；这样，每个 S_i 就是 H 的真值。反之，已知 $Pr_1({}^1p_1^1, {}^1p_2^1, \cdots, {}^1p_p^1, z_1, s)$，其中不包含除了它的 Arg 中变量以外的自由变量，那么如果存在一个网 H 和 H 中的一系列 N_i，使得 $N_1(z_1) \equiv . Pr_1(N_1, N_2, \cdots, z_1, sa_1)$是 H 的真值，式中 sa_1 的形式是 $N(0)$，我们就说它是**狭义可实现的**。如果对某个 n，$S^n(Pr_1)(p_1, \cdots, p_p, z_1, s)$ 在上述意义下是可实现的，我们就称它是**广义可实现的**，或简称**可实现的**。这里 c_{pi}是起实现作用的神经元。有两个神经刺激定律，如果根据一个假定，每一个 S 不管狭义广义都是可实现的，那么根据另一假定它也是可实现的，不过可能通过不同的网实现，这时我们将说这两个定律是两个等价的假定。

下面有关可实现性的定理都是指广义的。在某些情况下，可以得出较鲜明的有关狭义可实现性的定理；但是这除了

增加叙述的复杂性而外，很少有实际价值，因为我们目前的神经生理学知识所确定的刺激规律只是广义等价的，而更精确的定理会因我们采用的可允许的假定而有所不同。然而，不太精确的定理在等价的条件下是不变的，只要对冲动通过整个网的准确时间没有严格要求，所有的意图都能充分得到满足。

现在可以准确地陈述我们的中心问题了：第一，找到一个获取一组可计算的 S 的有效方法，该 S 构成任何已知网的解；第二，用一个有效的方式来表征这类可实现的 S。从实质上来说，这些问题是要演算任一网的行为，并且要找出一个按规定方式完成行为的网，如果这种网存在的话。

如果一个网包含一个环，也就是说，如果网上存在着一个神经元的链 c_i，c_{i+1}，…，链上每一个元通过突触与下一个元相连，首尾都一样，则这个网就被称为**循环**的。如果有一组这样的神经元 c_1，c_2，…，c_p，把它从 H 中移去，剩下的 H 就变成无环的，并且更小的神经元类别再没有这种特性，那么这组神经元就被称为**循环**组，并且它的基数是 H 的**阶**。我们将要看到，具有重要意义的是，一个网的阶是它行为复杂性的标志。特别是，零阶网具有非常简单的特性，我们首先来讨论它们。

让我们用下面的递归式来定义一个指称**时间命题函项**（**TPF**）的**时间命题表达式**（一个 TPE）：

1. ${}^1p^1[z_1]$是一个 TPE，其中 p_1 是谓项变量。

2. 如果 S_1 和 S_2 是包含相同自由个体变量的 TPE，那么 SS_1，$S_1 \vee S_2$，$S_1 . S_2$ 和 $S_i . \sim S_2$ 也是如此。

3. 除了(1)(2)外都不是 TPE。

定　理　Ⅰ

每一个零阶网都能用时间命题表达式来求解。

设 c_i 是 H 中阈值 $\theta_i > 0$ 的任一神经元，并设 c_{i1}，c_{i2}，…，c_{ip} 上分别具有刺激性突触 n_{i1}，n_{i2}，…，n_{ip}。设 c_{j1}，c_{j2}，…，c_{jq} 上有抑制性突触。设 k_i 是一组子类 $\{n_{i1}, n_{i2}, \cdots, n_{ip}\}$，它们的各元之和超过 θ_i。于是据上述假定，我们就能得到

$$N_i(z_1). = . S\left\{\prod_{m=1}^{q} \sim N_{jm}(z_1).\sum_{\alpha\epsilon k i}\prod_{\delta\epsilon\alpha} N_{is}(z_1)\right\} \qquad (1)$$

"$\sum$"和"$\prod$"是析取和合取的句法符号，在每一情况下它们都是有限的。因为对于每一个不是周围传入神经的 c_i 都能写出这种形式的表达式，所以用(1)中对应的表达式替换每一个不是周围传入神经的神经元的 N_{jm} 或 N_{is}，并且对该结果重复这个过程，最后就可以只用周围传入神经 N 得出 N_i 的表达式，因为 H 是无环的。此外，这个表达式是一个 TPE，因为(1)显然是 TPE；同时从定义可直接得出，用 TPE 替换 TPE 中的一个成分 $p(z)$ 得到的也是一个 TPE。

定　理　Ⅱ

每一个 TPE 都可用一个零阶网实现。

函子 S 显然可以用析取、合取和否定进行交换。显然，以任一狭义可实现的 S_i 替换一个可实现的表达式 S_1 中的 $p(z)$，其结果本身也是狭义可实现的；构造这种起实现作用的网的方法，是用那些在 S_i 的网中起实现作用的神经元替代 S_1 的网中的周围传入神经。如果 S_2 能够在狭义上得以实现，一个神经元网则在狭义上实现了 $p_1(z_1)$，图 1a 示出一

个网在狭义上实现了 $Sp_1(z_1)$,并且因而 SS_2 也是狭义可实现的。如果 S_2 和 S_3 是可实现的,那么对适当的 m 和 n, S^mS_2 和 S^nS_3 也是狭义可实现的。从而可以推及 $S^{m+n}S_2$ 和 $S^{m+n}S_3$。现在图 1b, c 和 d 中的网分别在狭义上实现了 $S(p_1(z_1) \vee p_2(z_1))$, $S(p_1(z_1).\, p_2(z_1))$ 和 $S(p_1(z_1).\, \sim p_2(z_1))$,从而 $S^{m+n+1}(S_1 \vee S_2)$, $S^{m+n+1}(S_1.\, S_2)$ 和 $S^{m+n+1}(S_1.\, \sim S_2)$也是狭义可实现的。所以如果 S_1 和 S_2 是可实现的,那么 $S_1 \vee S_2 S_1.\, S_2 S_1.\, \sim S_2$ 就是可实现的。经过完全归纳,所有 TPE 都是可实现的。由此可见,所有的网都能看作是由图 1a, b, c, d 这样的基础单元组合而成的,正像时间命题表达式是由进动、析取、合取,再联合否定的运算所产生的一样。特别是,在对一个网的状态,或对它的所有神经元的动作的真假值分布作出任何描述的情况下,除非这些值都是假的,总可以构成单个神经元,它的激发是使这描述有效的充要条件。此外,总是存在着一些数目不定的、拓朴结构不同的实现任何 TPE 的行为。

定 理 Ⅲ

设一个复杂语句 S_1 是已知的,它通过下列任何命题联系:否定、析取、合取、蕴涵和等价,以任何方式由一些 $p(z_1 - zz)$形式的基本语句组成,这里 zz 是任何数字,于是 S_1 就是一个 TPE,并且只有在假定它的成分 $p(z_1 - zz)$全部为假——亦即被假语句所替代时,它才为假,或是它的真值表的最后一行包含一个“假”,或是在只由否定项组成的它的希尔伯特析取的正则形式中,一个项也不存在。

这后三个条件当然是等价的(Hilbert and Ackermann

1938)。我们由归纳看到，第一个条件是必要的，因为当 $p(z_1 - zz)$被一个假语句替代时，$p(z_1 - zz)$就变成假的，而当 S_1 和 S_2 两个成分都为假时，$S_1 \vee S_2$，$S_1 . S_2$ 和 $S_1 . \sim S_2$ 全为假。注意到下述情况，我们就看到最后的一个条件是充分的：当一个析取的成分都是 TPE 时，它就是 TPE，同时任何项 $S_1 . S_2 . \cdots S_m . \sim S_{m+1} . \sim . \cdots \sim S_n$ 都能够写成 $(S_1 . S_2 . \cdots S_m) . \sim (S_{m+1} \vee S_{m+2} \vee . \cdots \vee S_n)$，它显然是一个 TPE。

有一些发生在过去的久远不定的事件，在条件说明中无从得到参考，对于这种情况来说，上述定理事实上就为构造这种要求下的神经网提供了一个十分方便和有效的方法。作为例子，我们可以考虑由瞬间冷引起发热感觉的情况。

如果一个冰冷的物体贴近皮肤片刻，然后移走，就会有热的感觉；如果作用的时间更长一些，则只有冷的感觉，而不再有开始时即使是瞬间的热感。我们知道，有一种皮肤感受器能够对热起反应，而另外一种能够对冷起反应。如果我们用 N_1 和 N_2 分别表示这两种感受器的动作，用 N_3 和 N_4 表示那些含有热冷感觉活动的神经元的动作，我们的要求就可写为：

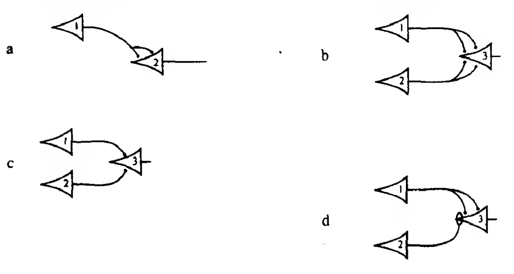


图 1-1

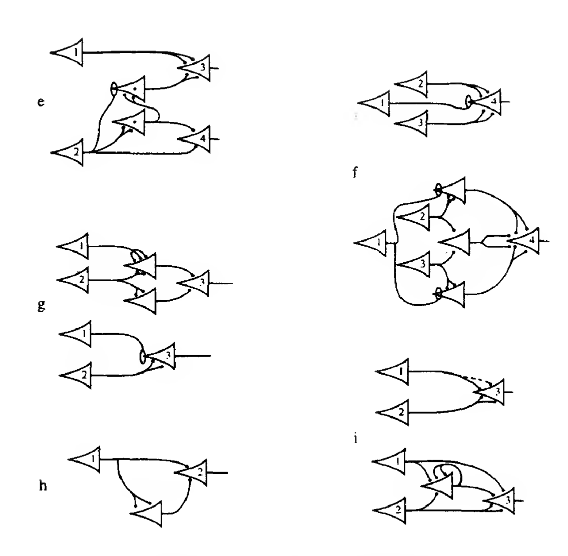


图 1-1 图中的表达式。图中神经元 c_i始终以细胞体上的数字 i 来标记,而对应动作则和正文中一样由带有下标 i 的“N”来表示。

a. $N_2(t) . \equiv . N_1(t-1)$

b. $N_3(t) . \equiv . N_1(t-1) \vee N_2(t-1)$

c. $N_3(t) . \equiv . N_1(t-1) . N_2(t-1)$

d. $N_3(t) . \equiv . N_1(t-1) . \sim N_2(t-1)$

e. $N_3(t) : \equiv : N_1(t-1) . \vee . N_2(t-3) . \sim N_2(t-2)$

$N_4(t) . \equiv . N_2(t-2) . N_2(t-1)$

f. $N_4(t) : \equiv : \sim N_1(t-1) . N_2(t-1) \vee N_3(t-1) . \vee . N_1(t-1) . N_3(t-1) . N_1(t-1)$

$N_4(t) : \equiv : \sim N_1(t-2) . N_2(t-2) \vee N_3(t-2) . \vee . N_1(t-2) . N_2(t-2) . N_3(t-2)$

g. $N_3(t) . \equiv . N_2(t-2) . \sim N_1(t-3)$

h. $N_2(t) . \equiv . N_1(t-1) . N_1(t-2)$

i. $N_3(t) : \equiv : N_2(t-1) . \vee . N_1(t-1) . \vee . N_1(t-1) . (Ex) t-1 . N_1(x) . N_2(x)$

$$N_3(t):\equiv:N_1(t-1).\vee.N_2(t-3).\sim N_2(t-2)$$
$$N_4(t).\equiv.N_2(t-2).N_2(t-1)$$

为简单起见，这里我们假定冷的感觉需要持续大约两个突触延迟，相比之下热的感觉只需一个延迟。这些条件显然低于定理Ⅲ。因此可用定理Ⅱ的方法构造出一个网来实现它们。开始时，我们用图 1-1a, b, c, d 中实现的运算，把它们写成能显示出它们组成成分的形式，即这种形式：

$$N_3(t).\equiv.S\{N_1(t)\vee S[(SN_2(t)).\sim N_2(t)]\}$$
$$N_4(t).\equiv.S\{[SN_2(t)].N_2(t)\}。$$

我们首先为括在最里面的函数构造一个网，然后向外发展；在这种情况下，我们运行一个由图 1-1a 中所示那种形式的网，比如说，从神经元 c_2 到某个神经元 c_a，得到结果：

$$N_a(t).\equiv.SN_2(t)。$$

下一步引入图 1-1c 和图 1-1d 形式的两个网，两者都从 c_a 和 c_2 开始运行，分别在 c_4 和比如说 c_b 结束。于是

$$N_4(t).\equiv.S[N_a(t).N_2(t)].\equiv.S[(SN_2(t)).N_2(t)]。$$
$$N_b(t).\equiv.S[N_a(t).\sim N_2(t)].$$
$$\equiv.S[(SN_2(t)).\sim N_2(t)]。$$

最后运行图 1-1b 形式的网，从 c_1 和 c_b 到 c_3，并导出

$$N_3(t).\equiv.S[N_1(t)\vee N_b(t)].$$
$$\equiv.S\{N_1(t)\vee S[(SN_2(t)).\sim N_2(t)]\}。$$

$N_3(t)$和 $N_4(t)$的这些表达式就是所要求的表达式；而这一起实现作用的网在图 1e 中完整地示出。

这种幻觉使得感知和"外部世界"之间的对应关系依赖于中介神经网特殊结构特性这一事实非常清楚。当然,对应于不同的网,采用其他各种有关皮肤感受器行为的假定,也能产生同样的幻觉。

现在我们来考虑某些等价定理,这些定理证明了神经刺激的各种替换定律,除时间而外,在本质上是等同的。我们首先来讨论**相对抑制**的情况。它意指这样一个假定:一个抑制性突触的激发并不绝对地阻止神经元的激发,而只是提高它的阈值,所以与别的情况相比,必须有更多的刺激性突触同时激发才能将它激发起来。我们可以不失一般性地假定,对每一个这样的突触激发,阈值的增长是1;于是我们就得到如下定理:

定　理　Ⅳ

相对抑制和绝对抑制在广义上是等价的。

根据(1)的方式,但改成运用相对抑制的假定,我们可以写出一个神经刺激定律来;经检查,这个表达式是一个 TPE。图 1-1f 给出一个用绝对抑制替代相对抑制的例子。相反的替代更为容易,我们分别给每个传入 c_i 的抑制性轴突以任何充分多数量的抑制性突触即可。

其次,我们考虑消退的情况。这可以写作神经元 c_i 激发后阈值 θ_i 的变化形式;对最接近的整数——只有对这样的近似而言,阈值变化在自然刺激形式中才是有意义的——来说,这可以写作激发后 j 个突触延迟的序列 $\theta_i + b_j$,其中当 j 足够大时,比如说 $j = M$ 或更大时,$b_j = 0$。这样我们就可以陈述定理 V。

定　理　Ⅴ

消退与绝对抑制是等价的。

假定相对抑制暂时保持，我们仅需运行 M 个环路 J_1，J_2，…，J_m，它们分别包含 1，2，…，M 个神经元，这就会使得任一环路上每个连接的激发都足以激发下一个连接，从神经元 c_i又返回到它自己，环路 J_j 的末尾恰好有 c_i上的 b_j 个抑制性突触。显然，这将产生所要求的结果。相反的替换可以由图 1－1g 中的模式来完成。从替代形式的传递性出发，我们就推导出这个定理。著名的定理Ⅵ也属于这组定理。

定　理　Ⅵ

助长和时间求和可以被空间求和替代。

显而易见，在受刺激细胞和要求在其上保持时间求和的神经元之间，我们只须引入一个增加突触数目的、适当的延迟链序列。这样，空间求和的假定就会给出所需要的结果。作为例子，请看图 1－1h。这个过程可用于表明，观察到的整个网的时间求和并不含有个体神经元之间相互作用中的那种机制。

学习现象具有在神经活动的大多数生理变化过程中保持不变的特性，似乎要求网的结构有可能出现永久性改变。这种改变的最简单的方式是形成新突触，或等价的局部阈值降低。我们设想某些轴突末梢开始时不能刺激后继的神经元；但如果在某一时刻这个神经元激发了，并且轴突末梢也同时受到刺激，末梢就变成通常类型的突触，从此以后就有了刺激

神经元的能力。抑制性突触的损失得出一个完全等价的结果。于是我们就有定理Ⅶ。

定 理 Ⅶ

可变的突触可用环来代换。

这一过程可用图 1 - 1*i* 中的方法来实现。还必须注意，一个变为自发活动的、并保持这一特性的神经元，同样也能由一个环来替代，该环进入活动状态靠的是活动开始时的周围传入神经，而它被抑制则是靠活动停止时的周围传入神经。

3. 理论：有环网

那些不满足以前不带环假定的网，处理起来要比满足的情况困难得多了。这主要是由于：活动可能在一个环路中形成，并在一段不确定的时间内继续环绕它震荡，所以可实现的 Pr 可能涉及对一些久远程度不定的过去事件的参照。我们来考虑这样一个网 H，假定它为 p 阶，并设 c_1，c_2，…，c_p 是 H 中的一组循环的神经元。首先从定义可知：H 的每一个 N_s 都可以表达为 N_1，N_2，…，N_p 和绝对传入神经的 TPE；于是 H 的解仅涉及对该循环组表达式的确定。照此办理，我们就能推导出一组表达式[A]：

$$N_i(z_1) . \equiv . Pr_i[S^{n_{i1}} N_1(z_1), S^{n_{i2}} N_2(z_1), \cdots, S^{n_{ip}} N_p(z_1)], \tag{2}$$

式中 Pr_i 也涉及周围传入神经。如果 n 是 n_{ij} 的最小公倍数，那么按照(2)式，以周围传入神经的等价物替换 N_j，并且在所得结果上将这个过程重复足够多的次数，我们就获得如下形式的 S：

$$N_i(z_1).\equiv.Pr_1[S^nN_1(z_1),\ S^nN_2(z_1),\ \cdots,\ S^nN_p(z_1)]。\tag{3}$$

这些表达式可以用希尔伯特析取的正则形式写成：

$$N_i(z_1).\equiv.\sum_{\substack{\alpha\epsilon k \\ \epsilon k \\ \beta\alpha}} S_a \prod_{j\epsilon k} S^nN_j(z_1) \prod_{j\epsilon\beta\alpha} \sim S^nN_j(z_1),$$

对适当的 k，　　(4)

式中 S_a 是一个网 H 的绝对传入神经的 TPE。这里大约有 2^p 个不同的语句，由 pN_i 构成，采用方法是把它们之中某一部分的合取与其余部分的否定的合取结合起来。若由 $X_1(z_1)$，$X_2(z_1)$，$\cdots$，$X_{2p}(z_1)$ 对其数值化，我们可以通过使用表达式(4)得出如下形式的一组等效方程：

$$X_i(z_1).\equiv.\sum_{j=1}^{2p} Pr_{ij}(z_1).S^nX_j(z_1)。\tag{5}$$

现在我们把下标数 i，j 引入对象语言，即定义 Pr_1 和 Pr_2，使得每当 zz_1 和 zz_2 分别代表 i 和 j 时，就可证明 $Pr_1(zz_1,\ z_1).\equiv.X_i(z_1)$ 和 $Pr_2(zz_1,\ zz_2,\ z_1).\equiv.Pr_{ij}(z_1)$。

这样我们可以把(5)式重写成：

$$(z_1)zz_p:Pr_1(z_1,z_3).\equiv.(Ez_2)zz_p.Pr_2(z_1,z_2,z_3-zz_n).Pr_1(z_2,z_3-zz_n)\tag{6}$$

式中 zz_n 代表 n，而 zz_p 代表 2^p。经过重复替代，对代表 s 的任何数字 zz_2 来说，我们就得到一个表达式：

$$(z_1)zz_p: Pr_1(z_1, zz_n zz_2). \equiv .(Ez_2)zz_p(Ez_3)zz_p\cdots(Ez_n)zz_p.$$
$$Pr_2(z_1, z_2, zz_n(zz_2-1)).Pr_2(z_2, z_3, zz_n(zz_2-1)).\cdots.$$
$$Pr_2(z_{n-1}, z_n, 0).Pr_1(z_n, 0)。 \quad (7)$$

通过归纳,不难证明它与下式等效:

$$(z_1)zz_p:.Pr_1(z_1, zz_n zz_2): \equiv :(Ef)(z_2)zz_2 - 1f(z_2 zz_n)$$
$$zz_p.f(zz_n zz_2) = z_1.Pr_2(f(zz_n(z_2+1)),$$
$$f(zz_n zz_2)).Pr_1(f(0),0) \quad (8)$$

既然这对所有 zz_2 都适用,因此下式也成立:

$$(z_4)(z_1)zz_p: Pr_1(z_1, z_4). \equiv .(Ef)(z_2)(z_4-1).f(z_2)$$
$$\leqslant zz_p.f(z_4) = z_1 f(z_4) = z_1.Pr_2[f(z_2+1), f(z_2), z_2].$$
$$Pr_1[f(res(z_4, zz_n)), res(z_4, zz_n)], \quad (9)$$

式中 zz_n 代表 n, $res(r,s)$ 是 r 除以 s 的余数,而 zz_p 代表 2^p。它可以近似地写成

$$N_i(t). \equiv .(E\Phi)(x)t-1.\ \Phi(x) \leqslant 2^p.\ \Phi(t)$$
$$= i.P[\Phi(x+1), \Phi(x).\ N\Phi(_0)(0)],$$

假定式中 x 和 t 也能被 n 除尽,而 Pr_2 代表 P。根据前面的论述,我们得到定理Ⅷ。

定 理 Ⅷ

表达式(9)表示网 H 的那组循环神经元,它和用这组神经元表达其他神经元动作的某个 TPE 一起,构成了网 H 的一个解答。

现在来考虑一组 S_i 的可实现性问题。由简单归纳可证得的第一个必要条件是,如果对 S_i 中其他自由变量 p 存在着

类似陈述，即任何神经网都不能计及未来的周围传入神经，则下式成立：

$$(z_2)z_1 . p_1(z_2) \equiv p_2(z_2) . \rightarrow . S_i \equiv S_i\begin{Bmatrix} p1 \\ p2 \end{Bmatrix} \qquad (10)$$

根据定义 $Pr_{mi} = \hat{f}[(z_1)(z_2)z_1(z_3)zz_p : . f(z_1, z_2, z_3)$

$$\equiv 0. \vee . f(z_1, z_2, z_3) = 1 : f(z_1, z_2, z_3)$$
$$= 1. \equiv . p_{z3}(z_2) : \rightarrow : S_i],$$

任何满足这个要求的 S_i 都能用如下形式的等效 S 来替代：

$$(Ef)(z_2)z_1(z_3)zz_p : f\epsilon Pr_{mi} : f(z_1, z_2, z_3) = 1. \equiv . p_{z3}(z_2) \qquad (11)$$

式中 zz_p 代表 p。现在考虑 α_i 类别的这些序列，对 α_i，

$$N_i(t) : \equiv : (E\Phi)(x)t(m)q : \Phi\epsilon\alpha_i : N_m(x).$$
$$\equiv . \Phi(t, x, m) = 1 \quad [i = q+1, \cdots, M] \qquad (12)$$

对某个网成立。这些将被称为**可理解**的类别。我们把一组 k 类别生成的**布尔环**定义为可通过重复应用逻辑运算由 k 的元形成的那些类别的集合；即

$$R(h)^{①} = p`\hat{\lambda}[(\alpha, \beta) : \alpha\epsilon k \rightarrow \alpha\epsilon\lambda : \alpha, \beta\epsilon\lambda.$$
$$\rightarrow . -\alpha, \alpha . \beta, \alpha \vee \beta\epsilon\lambda].$$

同时我们定义

$$\overline{R}(k). = . R(k) - \iota`p` - “k,$$
$$R_e(k) = p`\hat{\lambda}[(\alpha, \beta) : \alpha\epsilon k \rightarrow \alpha\epsilon\lambda.$$

---

① 此处和以下分别以“R”和“h”替代原文中书写体 R 和希腊文 k 的字型。——译者

$$\rightarrow . -\alpha, \alpha . \beta, \alpha \vee \beta, S``\alpha\varepsilon\hat{\lambda}$$

$$\overline{R}_e(k) = R_e(k) - \iota`p` - ``k,$$

以及 $\sigma(\Psi, t) = \Phi[(m).\hat{\Phi}(t+1, t, m) = \Psi(m)]$。

$R_e(k)$类别是仿照 $R(k)$由 k 形成的，但是不仅通过重复应用逻辑运算，而且也要重复应用由 $S(P)\varepsilon S``\alpha$ 替代特性 $P\varepsilon\alpha$ 类别的运算。于是我们得出**引理**：

$Pr_1(p_1, p_2, \cdots, p_m, z_1)$是一个 TPE 当且仅当

$$(Z_1)(p_1, \cdots, p_m)(Ep_{m+1}) : p_{m+1}\varepsilon \overline{R}_e(\{p_1, p_2, \cdots, p_m\})$$

$$p_{m+1}(z_1) \equiv Pr_1(p_1, p_2, \cdots, p_m, z_1) \qquad (13)$$

为真时；并且它是一个不包含"S"的 TPE，当且仅当"$\overline{R}_e$"被"$\overline{R}$"替代上式成立时，于是得到定理Ⅸ。

定 理 Ⅸ

一个类别 $\alpha_1, \alpha_2, \cdots, \alpha_s$ 的序列是一个可理解类别的序列，当且仅当

$$(Em)(En)(p)n(i)(\Psi) : . (x)m\Psi(x)$$

$$= 0 \vee \Psi(x) = 1 : \rightarrow : (E\beta)$$

$$(E\gamma)m . \Psi(\gamma) = 0 . \beta\varepsilon R[\hat{\gamma}((Ei) . \gamma = \alpha_i)] . \vee . (x)m .$$

$$\Psi(x) = 0 . \beta\varepsilon \overline{R}[\hat{\gamma}((E_i) . \gamma = \alpha_i)] : (t)(\Phi) : \Phi\varepsilon\alpha_i . \quad (14)$$

$$\sigma(\Phi, nt+p) . \rightarrow . (Ef) . f\varepsilon\beta . (w)m(x)t-1 .$$

$$\Phi(n(t+1)+p, nx+p, w) = f(nt+p, nx+p, w)。$$

这里的证明直接从引理得出。该条件是必要条件，因为每一个可以写成具有形式(4)表达式的网显然都能验证它，Ψ

是 S_a 的特征函数，而对每个 Ψ 来说，β 是其指称具有 $\prod_{i\varepsilon\alpha} Pr_i$ $\prod_{j\varepsilon\beta} Pr_j$ 形式的类别，此处对所有的 k，Pr_k 均代表 α_k。反过来，我们可以对一个实现了满足(14)式的可理解类别的网 H 写出具有形式(4)的表达式，方法是代入表示若干 Ψ 和一个 Pr 的 $Pr_a Pr$——仿照析取正则形式类别的方式写出，再代入表示与该 Ψ 对应的 α 的 $Pr_a Pr$，并与 Ψ 结合。由于具有形式(4)的每一个 S 是显然可以实现的，我们就得出这个定理。

我们能够根据当前的知识对各种特定网的全部过去状态确定到怎样的程度，即何时我们能构造一个网，要激活它那组循环神经元须要求周围传入神经具有一组由已知函数 Φ_i 所规定的过去的值，研究一下这种状况，是不无益处的。在这种情况下，上述定理的 α_i 类别简化为单元类别；而这一条件可以转换成：

$$
\begin{aligned}
&(Em,\ n)(p)n(i,\ \Psi)(Ej):.(x)m:\Psi(x)\\
&\quad =0.\vee.\Psi(x)=1:\\
&\Phi_j\varepsilon\sigma(\Psi,\ nt+p):\rightarrow:(w)m(x)t-1.\Phi_i(n(t+1)\\
&\quad +p,nx+p,w)=\Phi_j(nt+p,\ nx+p,\ w):.\\
&(u,\ v)(w)m.\Phi_i(n(u+1)+p,\ nu+p,\ w)\\
&\quad =\Phi_i(n(v+1)+p,\ nv+p,\ w)。
\end{aligned}
$$

由于篇幅的限制，我们对上述论点的介绍是十分简要的；我们打算在今后出版中扩充它和它的某些内涵。

最后一个定理的条件虽然不详细，但却简单明了；然而它在实际场合应用时，需要对大约 2^{2^n} 个函数类别，也就是对 $R(\{\alpha_1,\ \cdots,\ \alpha_s\})$的各个元进行考察。因为其中每一个都是

定理Ⅸ的一个可能的 β,这种结果不可能是清晰分明的。但是我们有可能得到一个 S 的可实现性的充分条件,这条件应用起来很方便,并且有可能适合于大多数实际用途。这由定理Ⅹ给出。

定　理　Ⅹ

让我们用下面的递归方式定义 S 中的一组 K:

1. 任何 TPE,以及任何自变量已由 K 中的元替代的 TPE,都属于 K;

2. 如果 $Pr_1(z_1)$是 K 的一个元,那么$(z_2)z_1.Pr_1(z_2)$,$(Ez_2)z_1.Pr_1(z_2)$和 $C_{mn}(z_1).s$ 都属于它,式中 C_{mn} 代表以 n 为模与 m 同余的特性,$m<n$。

3. 这组 K 不再有更多的元。

因此 K 的每个元都是可实现的。

因为如果 $Pr_1(z_1)$是可实现的,而

$$N_i(z_1).\equiv.Pr_1(z_1).SN_i(z_1)$$

$$N_i(z_1).\equiv.Pr_1(z_1)\vee SN_i(z_1)$$

是方程(4)的表达式的神经网,那么就分别实现$(z_2)z_1.Pr_1(z_2)$和$(Ez_2)z_1.Pr_1(z_2)$;具有 n 个连接、其中每一个都足以刺激下一个连接的简单环路 $c_1, c_2, \cdots, c_n$,给出了作为最后形式的表达式

$$N_m(z_1).\equiv.N_1(0).C_{mn}。$$

通过归纳,我们推导出这个定理。

最后,还要提到一件事。不难看出,首先,如果每个网配备一条传送带、一些与传入神经相联系的扫描器和适合于完

成必要运动运算的传出神经,那么它仅能计算像图灵机所能计算的那种数;第二,每一个后面的数都能由这样的网计算;而带环的网也都能由这样的网计算;即使在没有扫描器和纸带的情况下,带环的网也能够计算机器所能计算的某些数,但不能计算另一些,而并非所有的网都如此。为图灵的可计算性定义和它的一些等价物提供心理学证明是件有意义的工作,这些等价物包括丘奇的 λ 可定义性和克利尼的原始递归性:如果任何数可由一个有机体来计算,那么在这些定义下,它就是可计算的,反之亦然。

4. 结　论

因果关系必须具有状态描述与状态之间的必要联系规律,它以多种形式出现在多个学科中,但是除统计学而外,还从未像在本理论中一样是不可逆的。对任何一次传入神经刺激作用和任何一次所有作为构成成分的神经元的活动(每个“全或无”事件)所作的详细说明,确定了这一状态。对这些神经网所作的详细说明提供了必要联系规律,因而人们能从任一状态的描述中计算出对于后续状态的描述,但是析取关系包含的内容使得以前的状态不能完全确定。此外,作为构成成分的环的再生活动为过去时间提供的参照是不确定的。因此我们关于世界的知识,包括关于我们自己的知识,在空间上是不完全的,在时间上是不确定的。所有隐含在人类大脑中的无知,恰是那个使我们的知识变得有用的抽象性的互补物。在确定理论对观察和理论

对事实的认识关系中，大脑的作用真是再明白不过了，因为很显然，每一思想和每一感觉都是靠那个网之中的活动实现的，而任何这种活动都无法使实际传入神经得到完全确定。

在网发生变化时将其对事实的旧的、不完善的参照全部保留下来，能这样做的理论我们可能一个也没有，这样的观察我们也无法做到。耳鸣、刺痛、幻觉、妄想、慌乱和定向障碍都形成干扰。于是经验进一步证实，如果我们的神经网是不确定的，那么我们的事实就是不确定的，所以我们甚至连把一个性质或“形式”归属于“真实”的东西都做不到。随着网的确定，那个不可知的知识对象，即“自在之物”，就不再是不可知的了。

对心理学而言，无论它是如何定义的，有关网的详细说明都会为这一领域所能取得的所有结果作出贡献，即使这一分析被推向终极的心理单元——“心理元”，因为心理元恰恰就是单个神经元的活动。由于这种活动内在地带有命题的特点，所以一切心理事件都具有意向的或“符号的”特征。这些活动的“全或无”定律，以及它们的关系与逻辑命题关系的一致性，保证了心理元关系就是二值逻辑命题的关系。因此，在内省的、行为的或生理的心理学中，基本关系就是二值逻辑关系。

因此产生了整体论问题的构造性解，而整体论问题涉及分化的感觉认识连续统以及感知与执行的规范性、完善性和分解性。由因果关系的不可逆性可见，即使网是已知的，我们可从当前的活动预示未来，可是我们既不能从中枢神经系统活动推导出传入神经活动，也不能从传出神经活

动推导出中枢神经系统的活动，或是从当前的活动推导出过去的活动，这一结论因以下事实而显得更加肯定：目击者提供互相矛盾的证据，在诊断中难以区分有器质性病变的病人、歇斯底里病人和装病者，以及将一个人自己的记忆或回忆与他当时的记录进行对照。此外，一些系统可对再生网的传入神经与这一网中的特定活动这两者之间的差别作出响应，以至于使这一差别有所减少，从而表现出目的性的行为；我们知道有机体具有许多这样的系统，它们能促进体内平衡、嗜欲和注意力。因而，我们习惯于称为**精神**的那种活动的形式方面和终结方面都可根据当今的神经生理学严密地推导出来。与因果关系有关的明显结论，可使精神病学者得到安慰，这结论就是：对预测来说，历史决非是必要的。然而精神病学者观察到的现象只有根据神经活动才能得到解释，而这种神经活动至今仍超出他们的知识范围，这一同样有效的结论对他们则无甚用处。造成这一无知的关键是，从外表行为的任一样本到神经网的推理，在可构想出的网中并不是唯一的，然而实际存在的只有一个，并且随时可能表现出某种未曾预期到的活动。当然，对精神病学者来说，更主要的一点是：在这种系统中"心灵"不再"比鬼魂更可怕"。相反，病态心理可以不失范围或不失严谨地通过神经生理学的科学术语得到理解。就神经病学而言，这一理论明确区分了对已知活动来说是必要的网还是仅为充分的网，从而澄清了受干扰的结构与受干扰的功能的关系。在属于神经病学的领域中，等价的网与狭义等价的网之间的差别表明了神经活动时态研究的适用性和重要性：该理论为数学生物物理学提供了一个工具——用严格的符号方

式处理已知的网;同时也提供了一个简便的方法——根据需要的特性构造假设的网。

参考书目

Carnap, R. (1938). *The Logical Syntax of Language*. New York: Harcourt, Brace and Company.

Hilbert, D., and Ackermann, W. (1927). *Grunduge der Theoretischen Logik*. Berlin: J. Springer.

Whitehead, A. N., and Russell, B. (1925–7). *Principia Mathematica*. Cambridge: Cambridge University Press.



计算机器与智能

A·M·图灵*

1. 模仿游戏

我建议来考虑这个问题:“机器能够思维吗?”这可以从定义“机器”和“思维”这两个词条的涵义开始，定义应尽可能地反映这两个词的常规用法，然而这种态度是危险的。如果想通过检验它们通常是怎样使用的，从而找出“机器”和“思维”的词义，就很难避免这样的结论：这些词义和对“机器能够思维吗?”这个问题的回答，可以用类似盖洛普民意测验那样的统计学调查来寻找。但这是荒唐的。与这种寻求定义的做法不同，我将用另一个问题来替代这个问题，用作替代的问题与它密切相关，并且是用没有歧义的语言来表达的。

这个问题的新形式可通过一个游戏来描述，我们称之为“模仿游戏”。游戏由三个人来做，一个男人(A)，一个女人(B)，还有一个提问者(C)，性别不限。提问者待在一间与另两人分开的房子里。提问者在游戏中的目标是，确定另外

两人中哪一个是男性，哪一个是女性。他以标号 X 和 Y 称呼他们，在游戏结束时，他可能说“X 是 A，Y 是 B”，也可能说“X 是 B，Y 是 A”。提问者可以向 A 和 B 提出这样的问题：

C：X,请你告诉我你的头发长度,可以吗?

假定 X 实际上是 A,那么 A 必须做出回答。A 在游戏中的目标是,尽量使 C 作出错误判断。于是,他可能回答说:“我的头发是瓦盖式的短发①,最长的一束大约长 9 英寸。”

为了不让提问者从声调中得到帮助，这些回答应当写出来，若能打印出来则更好。理想的安排是，在两间房子之间，用一台电传打印机进行交流。也可以用一个中介人来重复提问和回答。第三个游戏参与者（B）的目标是帮助提问者。对她来说，最好的策略或许就是如实回答。她在回答时，可以加上这样的话：“我是女性，别听他的!”但是这种做法无济于事，因为那个男士也可以运用同样的方式。

现在我们要问的是:“如果在这个游戏中用一台机器代替 A,会出现什么情况?”在这种情况下做游戏时,提问者作出错误判断的次数,和他同一个男人和一个女人做这一游戏时一样多吗?这些问题替代了原来的问题:“机器能够思维吗?”

---

* A·M·图灵,“计算机器与智能”,选自《心灵》LIX, no. 2236(1950. 10):第 433—460 页。牛津大学出版社允许重印。

A·M·图灵(Alan M. Turing)英国数学家、逻辑学家。

① 该发型是女式的。——译者

2. 新问题的评论

我们除了问“什么是对新形式问题的回答?”还可以问“这个新问题值得去研究吗?”后一问题研究起来更为直截了当,从而可以制止无限回归。

新问题有利于在人的体力能力和智力能力之间划出一条截然分明的界线。任何工程师或化学家都认为,我们不可能制造出与人类皮肤一模一样的材料。即使到了某个时候,有可能做到这一点,但是假定这项发明成为现实,我们还是会觉得,试图用这种人工血肉把一台“思维机”装扮起来使它更像人类,是没有什么意义的。我们设定的这问题的形式,在防止提问者看到或接触到其他竞赛者或是听到他们的声音的条件下,反映了这个事实。这个判据的其他一些优点,可以在提问和回答的样本中得到说明。我们来看:

问:请以福斯河大桥为主题,给我写一首十四行诗。

答:这件事我可干不了,我从来不会写诗。

问:把 34957 与 70764 相加。

答:(停顿约 30 秒,然后给出答案)105621。

问:你会下象棋吗?

答:会。

问:我在我方的 K1 处有 K,再没有别的子了,你只剩 K6 处的 K 和 R1 处的 R,该你走了,你走什么呢?

答:(停顿 15 秒之后)R 到 R8 处,将死。

看来,问答法适合于引入几乎任何一个我们希望涉及的

人类需要花费心力的领域。我既不希望贬低不能在选美竞赛中有出色表现的机器,也不希望贬低同飞机赛跑失败的人。这个游戏的条件使这些能力缺陷成为不相干的。那些“证人”可以尽量夸耀他们的魅力、力量或英雄气概,如果他们认为这样做是可取的话,但是提问者不会要求这些实际方面的证明。

这个游戏可能会因为条件对机器太不利而招致批评。如果有人想要装作机器,他显然只能做出拙劣的表现。他会因做算术时的迟缓和不准确而立即暴露出来。但是,机器在做那些理应看成是思维的事情时,难道不会采用完全不同于人类的方式吗？这一反问是相当有力的,但是我们至少可以说,尽管如此,如果能够制造出一台机器,使它在模仿游戏中做出令人满意的表演,我们就不必再为这个反问操心了。

也许有人认为,在参与“模仿游戏”时,对机器来说,最好的策略可能并不是模仿人的行为。可能是这样的,但是我觉得这好像没有太大关系。不管怎样,这里不是打算研究游戏理论,所以我们可以假定,最好的策略就是设法提供那些人类会自然而然得出的答案。

3. 参与游戏的机器

如果我们不详细说明“机器”一词的意义是什么,我们在§1中提出的问题就不是很确切的。很自然,我们希望每一种工程技术都可以用于我们的机器;我们也希望允许这种可能性:工程师或工程师小组制造出一台能工作的机器,但是机器的制造者不能对它的运算方式作出令人满意的描述,因为

他们采用的方法主要是试验性的;最后,我们希望以正常方式出生的人不在机器之列。要框出一个能够满足这三个条件的定义,不是件容易的事。例如,有人可能强调工程师小组内的所有成员应当是同一性别的,但实际上这不会令人满意,因为还有从一个人的(比如说)皮肤单细胞里培养出一个完整个体的可能性。如果做到这一点,那是应给予最高奖赏的生物学技术的奇迹,但是我们并不会把它看作是"制造出思维机器"的例子。这促使我们放弃了每一种技术都可以被接受的要求。本文关于"思维机器"的兴趣是由一种通常称为"电子计算机"或"数字计算机"的特殊机器引发的,鉴于这个事实,我们更应该放弃上述要求。根据这一建议,我们只允许数字计算机参与我们的游戏。

初看起来,这个限制显得十分苛刻。我将证明,在现实中并非如此。要做到这一点,就需要对这些计算机的本质和特性作一简短说明。

或许有这样的说法:用数字计算机作为对机器的规定,就像我们有关"思维"的判据一样,如果数字计算机在这个游戏中不能做出良好的表现(这是与我们的信念相抵触的),这种规定就徒然令人不满意而已。

已有大量的数字计算机进入工作行列,所以我们可以问,"为什么不直接做实验呢? 它们也许很容易满足游戏的条件。可以利用大量的提问者去得出统计资料,以说明作出正确确认的情况有多少。"简单的回答是,我们并不是在问是否所有的数字计算机在游戏中都能够干得很出色,也不是在问当前可用的计算机是否能够干得出色,而是在问是否存在可以想象得到的能够干得出色的计算机。但这只是一个简单的回

答，稍后，我们将从不同的角度来思考这一问题。

4. 数字计算机

有关数字计算机的思想，可以从这种说法中得到解释：这种机器将完成人类计算机所能完成的任何运算；我们假定人类计算机是遵循固定规则的，他没有任何权力稍许偏离这些规则。我们可以假定这些规则是由一本书提供的，人一旦从事新的工作，这本书就要更换。还有无限多的纸张可供演算。他也可以在“台式机”上做乘法和加法，不过这并不重要。

如果我们把上述解释作为定义，就有进入论证循环的危险。为了避免这一点，这里提出一个可以取得满意效果的方法纲领。通常认为一台数字计算机是由三个部分组成的：

1）存储器

2）执行单元

3）控制器

存储器是存储信息的，相当于人类计算机使用的纸张，可以在这些纸上做演算，也可以用它来印出充满规则的书。当人类计算机在他的头脑里做演算时，一部分存储由他的记忆承担。

执行单元的作用是完成演算中所包含的各种具体的运算。这些具体运算的内容因机器而异，一般情况下，可以完成相当长的运算，如“3540675445 乘以 7076345687”，但是在有些机器中，只能做非常简单的运算，如“写出 0”。

上面提到过，供计算机使用的“规则书”可以用机器中的一部分存储替代，这时就称它为“指令表”。控制器的职责就

是监督这些指令按正常顺序正确执行。控制器的构造方式决定了这种结果必然出现。

存储器中的信息一般被分成大小适中的数据包。例如，在一台机器中，一个数据包可能由十个十进制数字构成。存储器中存放着各种信息的数据包，数字按照某种系统方式被分配到存储器的这些部分中。一个典型的指令可能指示道：

“把存储在 6809 位置上的数字加到存储在 4302 位置的数字上去，并把结果放入后一存储位置。”

当然，它不可能以英语表达方式出现在机器中，它很有可能以这种方式编码：6809430217。这里，17 的意思是指，在各种可能的运算中，要对这两个数字执行哪一种运算。在这一情况下，运算是如上所述：“把数字……加上。”我们注意到，这个指令用了十个数字，从而构成一个信息数据包，十分方便。在正常情况下，控制器会使这些指令按照它们存储位置的顺序一一执行，但是偶然也会遇到这样的指令：

“现在执行位于 5606 处的指令，然后由此继续。”

或者：

“如果位置 4505 存的是 0，接着执行存在 6707 处的指令，否则的话，按原有方式进行。”

上面这两种类型的指令非常重要，因为它们能够使一个运算序列一遍遍地被替代，直到某个条件满足为止，但是在这样做的过程中，不是每次执行新的指令，而是一遍遍地重复执行同一些指令。我们用一个家务例子作比喻：假定妈妈要托米每天早晨在上学的路上去鞋匠那里看一看她的鞋做好没有，她可以每天早晨向托米说一遍，她也可以用另一种一劳永逸的方法，即在厅里贴一张提示条，托米离家去上学时就能看

到，内容是叫他去鞋匠那里看看，并在他取到鞋之后把字条取去。

读者必须承认这个事实：数字计算机是可以按照我们所描述的原理制造的，而且的确已经制成了，同时，它们确实能非常接近地模仿人类计算机的行动。

说我们人类计算机是在使用充满规则的书，这当然只是一种方便的说法，真实的情况是，实际的人类计算机记住了他们要做的事情。如果要使一台机器模仿人类计算机在某种复杂运算中的行为，必须问一问人是怎样做到这一点的，然后把答案翻译成指令表形式。建立指令表通常被说成是“编程序”。“给机器编程序，让它完成运算 A”就意味着把一个恰当的指令表放入机器，使它执行 A。

在数字计算机思想的基础上出现了一个有趣的派生物，就是“带有随机元件的数字计算机”。它们具有像掷骰子这样的指令，或是某种等价的电子过程；比如说，这种指令有可能是“掷骰子，并把所得的数字放入存储位置 1000”。有时这样的机器被说成是有自由意志的（虽然我自己并不使用这个词）。一般情况下，不大可能通过观察机器来确定它是否带有随机元件，因为像根据 π 的十进制小数数字进行选择的装置，也能产生出类似的结果。

大多数实际的数字计算机，存储都是有限的。使计算机带有无限存储的思想，在理论上不存在困难。当然，在任一时刻，所使用的部分只能是有限的。同样，所构成的内容也只能是有限的，但是我们可以设想根据需要增加越来越多的内容。这样的计算机具有特殊的理论意义，可以称为无限容量计算机。

数字计算机的思想是一个古老的思想。1828—1839年间任剑桥大学卢卡斯讲座的数学教授巴比奇(C. Babbage)设计了一台这样的机器,叫作分析机,但是始终没有完成。尽管巴比奇已经掌握了所有的基本概念,但是在那个时代,他的机器没有表现出那种十分诱人的前景。它可能达到的速度,肯定要比人类计算机快一些,但是比曼彻斯特机慢大约100倍,而在现代机器中,曼彻斯特机本身也是比较慢的一种。巴比奇设计的存储完全是机械的,使用轮子和卡片。

巴比奇的分析机完全是机械的这一事实,有助于我们摆脱一个迷信。现代数字计算机是电子的,神经系统也是电子的,人们往往很看重这个事实。既然巴比奇的机器不是电子的,同时所有数字计算机在某种意义上又是等价的,因此我们了解到使用电这一点在理论上不可能很重要。当然,在信号快速传输的地方,总是离不开电的,所以我们在这两个方面发现它,也不足为奇。在神经系统中,化学现象至少和电一样重要。在某些计算机中,存储系统主要是声音的。因此,使用电这一特点看来只是表面上的相似性。我们与其寻求这样的相似性,还不如寻求数学上的功能相似性。

5. 数字计算机的普适性

上一节提到的数字计算机,可以归入"离散状态机"一类,这些机器的运动是通过突然跳动或是通过棘轮,从一个完全确定的状态,转变到另一个完全确定的状态。这些状态截然不同,所以它们之间不存在混淆的可能。严格地说,这样的

机器并不存在,每一物体的真实运动都是连续的。但是对许多类型的机器来说,把它们看作是离散状态机,更为有利。例如,考虑照明系统的开关时,我们假定开关的每一位置是绝对的开或绝对的关,就会很方便。中间位置肯定是存在的,但在大多数情况下,可以忽略它们。我们以一个轮子为例,来说明离散状态机。轮子的棘轮每秒钟转动一次,转 120 度,但是外部有一个操作杠杆,可以使它停止;同时,当轮子转到某个位置时,一盏灯就会亮起来。这个机器可用如下抽象方式来描述:机器的内部状态(用轮子的位置来描述)记为 q_1、q_2、q_3;输入信号是 i_0 或 i_1(杠杆的位置);任一时刻的内部状态可根据下表,由它的终端状态和输入信号来确定。

终端状态

		q_1	q_2	q_3
输入	i_0	q_2	q_3	q_1
	i_1	q_1	q_2	q_3

输出信号,作为内部状态唯一的外部可见指标(灯光),可由下表描述:

状态	q_1	q_2	q_3
输出	o_0	o_0	o_1

这是离散状态机的一个典型例子。若是离散状态机的可能状态数只是有限的,就可以通过这种表来描述它们。

看来,一旦确定机器的初始状态和输入信号,就总是可以预见所有未来的状态。这使我们想起拉普拉斯的观点:从宇宙在某一时刻由它的所有粒子的位置和速度所描述的完整状态出发,就可以预言出所有的未来状态。然而,同拉普拉斯所作的预言相比,我们拟议中的预言更具可行性。“宇宙整体”

系统有这种特点:初始条件中微小的错误,在其后的时间中,会产生出势不可挡的效果。一个单个的电子在某一时刻发生十亿分之一厘米的位移,就可能造成一个人一年后被雪崩杀死还是死里逃生这样重大的区别。这种现象不会在我们称之为"离散状态机"的机械系统中发生,这是该系统的基本特性。即使我们考虑的是实际的物质机器,而不是理想化的机器,在某一时刻对于某一状态的足够准确的知识,在其后任何数量的步骤上,仍能产生出足够准确的知识。

如前所述,数字计算机属于离散状态的机器类别。但是这种机器所能有的状态数一般是相当大的。例如,现在在曼彻斯特工作的那台机器,其状态数大约是 $2^{165,000}$,也就是约 $10^{50,000}$。把它和上述例子中有三个状态的棘轮加以比较,不难看出其状态数为什么会如此巨大。计算机中有存储器,它相当于人类计算机使用的纸张。必须有可能把写在纸上的任何一种符号组合写入存储器。为简单起见,假定只使用从0到9的数字作为符号,各种手写体不予考虑。假定计算机可接受100张纸,每张纸有50行,每一行有30个字的位置。那么状态数就是 $10^{100\times50\times30}$,即 $10^{150,000}$。这约等于三台曼彻斯特机并在一起的状态数。状态数以2为底的对数通常称为机器的"存储容量"。这样曼彻斯特机的存储容量大约是165,000,而我们例子中的那个轮机的容量大约是1.6。如果把两台机器并到一起,它们的容量必须相加,才能得到合成机的容量。这种情况可能导致这样的陈述:"曼彻斯特机含有64个磁道,每一磁道的容量是2560,还有8个电子管,每个容量为1280,其他各项的存储量大约是300,这样总计就是174,380。"

一旦有了和离散状态的机器相对应的表,就可以预言它将要做什么。完全有理由说,可以用数字计算机来执行这样的演算。如果数字计算机干得足够快,它就能够模仿任何离散状态机的行为。那么,如果用当前的机器(作为B)和模仿数字计算机(作为A)来表演模仿游戏,提问者是无法把它们区分开来的。当然数字计算机不仅要工作得足够快,还要有足够的存储容量。此外,必须对每一台要进行模仿的新机器重新编程。

由于数字计算机具有能模仿任何离散状态的机器这一特殊性能,它们被说成是**万能**机器。存在着具有这种特殊性能的机器的意义是重大的:如果排除速度因素,就没有必要为完成各种不同的计算过程,而去设计各种不同的新机器。所有计算都可以用一台数字计算机来完成,在每一种情况下只要配以适合的程序即可。由此可以认为,在一定意义上,所有数字计算机都是等价的。

现在我们再来看看§3结尾处提出的那个观点。我们曾经试探性地提出"机器能够思维吗?"这个问题,可否代之以"存在着可以想象得到的能够在模仿游戏中干得出色的数字计算机吗?"如果我们愿意的话,我们可以使这问题看起来更具一般性,即问:"存在着能够干得出色的离散状态的机器吗?"但是从普适性的角度出发,我们认为上述两个问题都与下述问题等价:"我们只讨论一台特定的数字计算机C。如果经过改进,使这台计算机具备合乎要求的存储,动作速度也相应提高,同时还为它提供了恰当的程序,那么在模仿游戏中,由人担任B的角色,C就能够令人满意地扮演A的角色,这有可能成为事实吗?"

6. 对主要问题的反对意见

现在我们可以考察一下这个已经得到澄清的基础，并准备就我们的问题“机器能够思维吗？”以及上一节结束时提到的该问题的转换形式进行辩论。我们不能完全放弃这个问题的原有形式，因为在替换形式恰当与否的问题上，存在不同见解，我们至少应该听一听与之有关的意见。

如果我先就这一问题谈谈我自己的意见，对读者来说，情况可能会更加明了。首先来看一下这一问题的较精确的形式。我认为，在大约50年的时间里，有可能对具有约10^9存储容量的计算机进行编程，使得它们在演示模仿游戏时达到这样出色的程度：经过5分钟提问，一般提问者作出正确判断的机会，不会超过70%。我认为，原来的问题“机器能够思维吗？”意义太不明确，因而不值得讨论。但是我认为，到本世纪末，在使用词以及由教育形成的一般见解方面将会发生很大变化，使得人们在谈论机器思维的问题时，可以不再担心有矛盾发生。我还进一步认为，把这些意见隐蔽起来，对任何有用的目标都是不适宜的。有一种流行的观点认为，科学家坚持不懈地沿着从确认的事实到确认的事实的路线前进，而从不受任何有改进的猜想的影响，这观点是相当错误的。假如澄清了什么是已经证明的事实，什么是猜想，不会产生什么害处。猜想是极其重要的，因为它们可以为研究提供有用的思路。

下面就来看看那些与我的见解相反的意见。

神学反对意见

思维是人类不朽灵魂的一种机能。上帝把不朽的灵魂给了每个男人和女人，而没有给任何其他动物和机器。所以任何动物和机器都不能思维。[①]

我完全不能接受这种说法，但是我愿意用神学方式来寻求回答。我觉得，把动物和人归入同一类，这样的论点更令人信服，因为在我的思想中，典型的有生命之物和无生命之物之间的区别，要比人和其他动物之间的区别更大些。如果我们考虑到其他宗教群体成员对待这个问题的态度，这一正统观点的武断性就会显得更加清楚。穆斯林认为女人没有灵魂，对于这个观点，基督徒是怎样想的呢？但是让我们撇开这一点不谈，言归正传。在我看来，上面引用的论点中隐含着对全能上帝无限权力的一种严重的限制。应该承认，有些事情上帝无法做到，比如让 1 等于 2，但是难道我们不应该相信上帝有这样的自由：他如果认为合适而把灵魂给予一头大象吗？我们或可认为，上帝只在结合着能给大象一个恰当改进的脑以满足灵魂需要的基因突变的情况下才运用这一权力。对机器来说，也可以作出形式上完全相似的论证。其不同之处也许是后者更难以"下咽"。但是这实际上只不过意味着，我们认为，上帝是不大会考虑哪些环境是适合于赋予灵魂的。这里谈到的情况将在本文的其他部分讨论。在打算构造这种机器的

① 这一观点也可能被指为异端。圣·托马斯·阿奎那〔《神学大全》，引自 B·罗素（Russell 1945: 458）〕指出，上帝不能不给人以灵魂，但这并不是对上帝权力的真正限制，而仅仅因为人的灵魂是不朽的，从而是无法毁灭的。

时候，我们不应当比我们在繁衍后代时更加不敬地僭越上帝创造灵魂的权力，不管就这两种情况的哪一种而言，我们都不过是上帝意志的工具，是为他所创造的心灵提供住所而已。

然而这只不过是一种臆想。至于神学论点，无论它们可以用来支持什么，都不会给我留下深刻印象。历史上，像这样无法使人满意的论点屡见不鲜。在伽利略时代，人们认为教科书上的"太阳仍站在那里……整个白天里它都不急于落下"(《约书亚记》10:13)和"上帝为地球安放了基础，任何时候它都不会移动"(《诗篇》105:5)是反驳哥白尼理论的充分理由。从今天的知识来看，这样的论点毫无价值。当一种知识失效之后，人们的印象就大不相同了。

"把头埋在沙中"的反对意见

"机器思维的后果太可怕了，我们希望并且相信机器做不到这一点。"

如此坦白的论点实属难得。但是大部分认真思考这一问题的人，都受到它的影响。我们愿意相信，人类以某种微妙的方式优于其他生物。如果能够证明人类**必然优越**，那是再好不过的，这样，人类就没有失去其居高临下地位的危险。神学论点的流行显然与这种情感有关。在知识阶层中，这种情感似乎十分强烈，因为他们比别的人更加看重思维的能力，并且更倾向于把他们有关人类优越性的信念建立在这种能力的基础上。

我认为这个论点没有多大价值，不值一驳，给一点安慰可能更恰当些，这种安慰也许应该从灵魂轮回说中去找。

数学上的反对意见

数理逻辑的许多结果都可以用来证明:离散状态机器的能力是有限度的。这些结果中最著名的就是所谓的哥德尔定理(Godel 1931),它表明,在任何功能充分的逻辑系统中,都可以形成一些陈述,它们在系统内部既不能被证明,也不能被证伪,除非是这种情况:系统本身不一致。还有一些在某些方面与之类似的结果,它们来自丘奇(Church 1936),克利恩(Kleene 1935),罗素和图灵(Turing 1937)。后一结果研究起来最方便,因为它直接涉及机器,而其他的只能用于较为间接的论证,例如使用哥德尔定理时,必须增加一些用机器描述逻辑系统以及用逻辑系统描述机器的方法。该结果所涉及的机器本质上属于有无限容量的数字计算机式的类型。这一结果认为存在着这种机器做不到的某些事情。如果像在模仿游戏中那样,使装备好的机器回答问题,就会出现这样的后果:对某些问题或是作出错误回答,或是根本无法回答,尽管回答的时间不受限制。这样的问题无疑有很多,这一台机器不能回答的问题,另一台机器也许能作出令人满意的回答。当然,我们暂且假定,问题属于只需作出"是"或"不是"这种回答的类型,而不是像"你对毕加索有什么看法?"这样的问题。我们知道这些机器肯定无法回答属于这种类型的问题:"假如对机器作如下详细说明……这机器对任何问题都回答'是'吗?"省略号将代之以对某一机器所作的标准形式的描述,有可能像§5中使用的那种形式。当所述机器与接受提问的机器具有某种比较简单的关系时,可以证明,或是作出错误回答,或是无法

回答。这是一个数学结论：经论证，证明机器是有能力缺陷的，而人类智能不比机器逊色。

对这个论点的简单答复是，虽然已经证明了任何一台特定机器的能力都是有限的，但在没有任何一种证据的情况下，所阐明的只是：对于人类智能，这种限度是不适用的。但是我认为这个观点不可能如此轻易地驳倒。只要向一台这样的机器提出恰当的难以回答的问题，当它给出一个确定的答案时，我们知道这个答案必然是错误的，这样我们就产生了某种优越感。这种感觉是虚幻的吗？它无疑是真实的，但是我认为不必过分看重这一点。我们因证明机器易犯错误而沾沾自喜，这是没有道理的，因为我们自己在回答问题时出的错是够多的了。进一步说，我们之所以产生优越感，是因为在和一台机器打交道时，取得了优于它的小小胜利。当然，也有可能是同时战胜所有机器。那么简言之，可能有一些人比任何现有的机器都聪明，但是另一方面，也可能有另一些比人更聪明的机器，如此等等。

据我看，那些持有数学论点的人，大都愿意接受模仿游戏作为讨论的基础。那些相信前两种反对意见的人，很可能对任何判据都不感兴趣。

有关意识的论点

杰斐逊教授在1949年的利斯特讲演中将这个论点表达得淋漓尽致，我引用其中一段：

“除非机器能够做到因为有思想、懂感情而不是通过符号的偶然来临去写十四行诗或创作协奏曲，否则我们

不能认为机器与大脑是等同的——也就是说，不仅把它写出来，而且知道已经把它写出来了。任何机器都不可能感觉到(不仅是人为地发出信号而已，这种设计是简易的)成功时的愉悦和电子管烧毁时的悲伤，也不会因听到奉承而兴奋，因犯错误而苦恼，因见到异性而着迷，在愿望实现不了时发怒或沮丧。”

看来，这个论点否定了我们试验的有效性。从这个观点的最极端的形式来看，确认一台机器能否思维的唯一办法，就是**变成**这台机器，并感受到自己在思维。然后可以向世人描述这些感受，但是，毫无疑问，任何人所做的任何介绍都不能被认为是正确的。同样，根据这一观点，得知**一个人**会思维的唯一方法，就是变成这个特定的人。这实际上是唯我论的观点。这也许是可能持有的最符合逻辑的观点，但是它使思想交流发生困难。A倾向于认为“A会思维，而B不会思维”，B却认为“B会思维，而A不会思维”。我们不再就这一观点继续争论，出于礼貌习惯，一般认为每个人都可以思维。

可以肯定，杰斐逊教授并不想采取极端的、唯我论的观点，他可能十分愿意接受模仿游戏作为试验。这个游戏(在没有参与者B的情况下)在实际运用时往往叫作口试，目的是了解某个人是真正懂得某个事物，还是“死记硬背”。让我们听一段这样的口试：

提问者：你的十四行诗的第一行为“让我把你比作夏日”，是不是同样可以用“春日”，甚至更好些？

参试者：这样不合韵律。

提问者：“冬日”怎么样？这完全合乎韵律。

参试者:是的,但是没有人愿意被比作冬日。

提问者:你是不是认为匹克威克先生让你想起圣诞节?

参试者:多少有一点。

提问者:但是圣诞节是一个冬日,我想匹克威克先生是不会反对这种比较的。

参试者:我认为你不够认真。冬日的意思是一个有冬天特征的日子,而不是一个像圣诞节那样的特殊的日子。

如此等等。如果写作十四行诗的机器在口试中能够作出这样的回答,杰斐逊教授会说什么呢?我不知道他是否还会认为机器的回答"只不过是人为地发出信号",但是如果回答能像上段中那样令人满意,并且滔滔不绝,我认为他就不会把机器说成是"一种简易的设计"。我想,这个短语所指的不过是这种装置:在机器中装入一段某个人阅读十四行诗的记录,再加上一个恰当的开关,可随时打开它。

总之,我认为可以说服大多数支持意识论观点的人放弃这一论点,而不是把他们推向唯我论的立场。这样的话,他们也许就愿意接受我们的试验。

我不希望给人以我认为意识与神秘性无关的印象。例如,某种悖论的内容就与把神秘性限制在一定范围内的企图有关。但是我认为,在我们能回答本文所涉及的那些问题之前,我们不一定非得将这些神秘性弄清楚。

有关各种能力缺陷的论点

这些论点是这样的:"就算你真的能够让一台机器完成你提到的所有那些事情,但是你决不可能让一台机器做到

X。"这里 X 的特征很多,下面只是其中的一部分:

要仁慈,机智,漂亮,友好,有首创精神,有幽默感,能辨别是非,会犯错误,会坠入情网,爱吃草莓冰淇淋,能让别人爱上自己,会从经验中学习,用词得当,能成为自己思想的主题,像人一样行为多变,做出某件全新的事情。

一般情况下,这些说法是无根据的。我认为,它们主要建立在科学归纳原理的基础上。人在一生中见过成千上万的机器,他从在机器身上所看到的东西中,得出许多一般性结论:它们又粗又笨,每台机器只能用于非常有限的目标,目标稍有变化,它们就变得毫无用处,无论何种机器,行为变化都非常之少,等等。他自然得出这样的结论:这些是机器一般的必要特性。在这些局限性中,许多与大多数机器存储容量过小有关。(我假定将存储容量的概念以某种方式扩展到离散状态机以外的那些机器上去。准确的定义并不重要,因为在当前的讨论中并不要求数学上的精确性。)几年前,很少听说过数字计算机,如果只谈它们的特点,而不讲它们的构造,那就很可能对它们产生许多疑惑。这大概是由于类似地运用了科学归纳原理,当然,这原理的这些运用在很大程度上是无意识的。挨过烧的孩子怕火,并且通过躲避火表现出他怕火,这时,我就可以说他是在运用科学归纳。(当然,我也能用许多别的方式来描述他的行为。)人类的工作和习惯看来不大适合作为科学归纳的素材。要获得可靠的结果,必须在相当大的时空范围中进行研究,否则,我们有可能(像许多英国儿童那样)判断说,人人都讲英语,学法语真傻。

然而,针对前面提到的许多能力缺陷,有一些专门评论。没有品尝草莓冰淇淋的能力,在读者看来,是无足轻重的。造

一台能品尝这种美味的机器并非不可能,但是试图使人们这样做,近乎白痴。重要的是,这种能力缺陷会引起另一些能力缺陷,比如难于在人与机器之间形成像白人和白人或黑人与黑人之间那样的友谊。

"机器不会出错"这种断言显得有点奇特,有人甚至反驳说"这样是不是更糟呢?"但我们还是采取较为赞同的态度,看一看它真正的意义是什么。我想,可以通过模仿游戏来解释这个批评。有一种看法是,提问者只要让人和机器做一些算术题,就可以把两者区分开来。机器会暴露,因为它太准确了。对此作答复并不困难。(为做游戏而配有程序的)机器不必力求对算术问题作出**正确**回答,它可以在演算方式中有意制造一些错误,来迷惑提问者。由于对算术错误类型选择不当,机器的失误也会自己暴露出来。即使这样来解释这一批评,其赞同的程度仍不够充分。但是我们进一步发展的余地已经不大了。在我看来,这种批评其实是混淆了两种类别的错误。我们可以称其为"功能性错误"和"结论性错误"。功能错误是由于有某种机械的或电的问题,使机器不能按照设计方式运行。在哲学讨论中,人们倾向于忽略这种错误的可能,因而所讨论的是"抽象机器"。这些抽象机器是数学构想,而不是物理对象。根据定义,它们不会造成功能性错误。在这个意义上,我们确实可以说"机器从不出错"。结论性错误只有在机器的输出信号具有某种意义时才会出现。例如机器可能打印出数学方程或英语句子。如果机器打出一个错误的命题,我们就说它犯了结论性错误。显然没有任何理由说机器不会造成这种错误。它也可能除了反复地打出"0 = 1"以外什么也不做。举一个较合常理的例子,它可能运用通过科学归

纳得出结论的方法。我们一定会估计到,这种方法偶尔也会产生错误的结果。

这种看法认为,机器不可能成为自己思想的主题。要对此作出回答,当然只有证明机器对**某种**题材具有**某个**思想。然而"机器运算的题材"看来必然意味着什么东西,至少对有关的人来说是如此。举例说,如果机器试图解方程 $x^2-4x-11=0$,就不妨认为,在这一刻这个方程就是机器的题材。在这种意义上,机器毫无疑问地可以是自己的题材。它可以用来协助机器编自己的程序,或是预测它自己结构变化后的结果。通过观察自己行为的结果,机器可以修改自己的程序,以便更有效地实现某种目标。这些都有可能在不久的将来实现,而不是乌托邦式的梦想。

那种批评机器不可能有丰富多采的行为的说法,等于在说机器不可能有丰富的存储容量。直至最近,甚至一千个数字的存储容量还是很罕见的。

我们这里提到的种种批评,常常以意识论点的形式伪装起来。通常只要知道了机器**能够**做出一件这样的事情,并对机器可能使用的那种方法加以描述,人们就不会依赖于印象了。大家都认为,方法(不管它可能是什么,因为它必然是机械的)的确是相当低级的。我们可以比较一下前面引用的杰斐逊讲演中括号内的文字。

洛夫莱斯夫人的反对意见

有关巴比奇分析机的最详细资料,来自洛夫莱斯夫人的回忆录(Lovelace 1842)。她在回忆录中写道:"分析机无权

说它创造出什么新的东西。它所能做的都是**那些我们知道怎样命令它**去执行的事情。”(着重号为她自己所加)哈特里(Hartree 1949)引用了这一陈述,并补充说:

> 这并不表明,没有可能建造一台这样的电子装置:它会“为自己着想”,或者用生物学术语说,人们能在其内部建立条件反射,以此作为“学习”的基础。这在原理上是否可能,是一个刺激人和令人兴奋的问题,某些新近的发展已经暗示出这一点。但是,看来那个时代制造或设计的机器并不具有这种特性。

在这一点上,我与哈特里完全一致。应当看到,他并不是认为所讨论的机器没有得到这种特性,而是认为洛夫莱斯夫人所能获得的证据没有使她相信机器具有这种特性。从某种意义上说,这种机器得到这一特性的可能非常之大,因为我们可以设想某一离散机器具有这种特性。分析机是一种普适的数字计算机,所以只要有相适应的存储容量和速率,就可以通过适合的程序,用来模仿我们所讨论的机器。伯爵夫人或巴比奇不大可能提出这种观点。无论如何,他们并没有责任断言所有能够断言的东西。

这个问题将在以“学习机”为题的一节中完整地加以考虑。

洛夫莱斯夫人反对意见的另一种说法是,机器“从来不能做任何全新的事情”。我们可以暂时用谚语来对答:“太阳底下没有新东西。”谁能断言他所做的“原创性工作”不过是由老师在他身上播下的种子生长而成的呢?或者不过是从众所周

知的一般原理中得到的结果。这种反对意见还有一种较好的说法:机器永远不会“使我们出乎意料”。这一说法是一种更直截了当的挑战,也可以直截了当地与之交锋。机器让我出乎意料的时候非常之多。这在很大程度上是因为我没有对于要确定期待机器做什么作出充分的演算,或者毋宁说是因为虽然我做过演算,但是演算得有点匆忙、马虎,有点冒险。我也许对自己说:“我假定这儿的电压应当和那儿的一样,反正这样假定就是了。”毫不奇怪,我会经常出错,得到出乎意料的结果,因为在实验做完时,我已经忘记了这个假定。这些自供会招致人们对我不良作风的谴责,但是人们不会怀疑我所说的意料之外体验的真实性。

我并不指望这个答复会使批评我的人保持缄默。他可能会说，那种出乎意料是由于我身上的某种创造性心理活动造成的，并不能为机器增光。这样，我们又被带回出自意识的论点，而远离出乎意料的这一概念。我们应该认为这种争论已经结束，但是也许值得指出，把某一事物评价为出乎意料的，同样需要“创造性心理活动”，无论出乎意料的事是源自一个人、一本书、一台机器，或是其他什么东西。

我相信，机器不可能让人出乎意料的观点起因于哲学家和数学家们特别容易持有的一个谬见。它假定：一旦事实呈现于心灵，这个事实的全部后果就会与之同时涌入心灵之中。在很多场合，这是一个很有用的假定，但是人们太容易忘掉它错误的一面。这样做的自然结果是人们就此假定，仅仅从数据和一般原理推出结论，是没有什么价值的。

有关神经系统连续性的论点

神经系统肯定不是一台离散状态的机器。神经脉冲冲击神经元时，有关它大小的信息方面出现一个小小的错误，就可能造成输出脉冲大小的巨大差别。可以认为，如果这样，就不能指望用离散状态系统模仿神经系统的行为。

诚然，离散状态的机器肯定不同于连续机，但是如果我们遵照模仿游戏的条件，提问者也无法从这种不同中得到什么好处。如果我们考虑另外一台较为简单的连续机，情况就会更加清楚。微分分析机就很合适。（微分分析机是用于某些演算的一种非离散状态型的机器。）有些微分分析机用打印方式给出答案，所以适合于做这个游戏。一台数字计算机虽然不可能精确地预言微分分析机怎样回答一个问题，但是它完全可以给出答案的正确类型。例如，如果需要求出 π 的值（实际值约为 3.1416），合理的做法是在 3.12，3.13，3.14，3.15，3.16 各值中任意选取，它们的概率分别是（比如说）0.05，0.15，0.55，0.19，0.06。在这种情况下，提问者很难把微分分析机与数字计算机区别开来。

有关行为的非形式特性的特点

建立一组规则，以说明一个人在所有情况下应该做什么，这是无法办到的事情。例如，有这样一个规则，看到红色交通灯时应该停下来，看到绿灯时可以行进，但是，若由于某种故障，两灯同时亮起来，该怎么办？人们可能会作出判断：停

下来最安全。但是某个新的难题又会随之而来。想要为所有的突发性事件提供指导规则，即使就交通灯的情况来说，也是办不到的。我完全同意这一点。

由此，就论证了我们不可能是机器。我想重新提出这个论点，只是担心有欠公允。也许可以采用这样的说法："如果每个人都有一套确定的制约他的生活的指导规则，人就并不比一台机器更强；但是这样的规则是不存在的，所以人不可能是机器。"这里最醒目的是不周延中项。我认为这个论点从未这样提出过，但是我相信这毕竟是一个曾经使用过的论点。然而，"指导规则"与"行为规律"之间可能有某种混淆，致使这个问题模糊不清。我所谓的"指导规则"指的是像"看到红灯即停"这样的条例，人们可以根据它来行动，也能意识到它的存在。而所谓"行为规律"则是指适用于人体的自然规律，例如"如果你刺痛他，他就会尖叫起来"。如果我们用"制约他的生活的行为规律"来替换上述论点中的"制约他的生活的指导规律"，那么不周延中项就不再是不可逾越的。因为我们相信，受到行为规律制约，就意味着是某种机器(虽然不一定是离散状态的机器)，这是正确的，不仅如此，反过来，是这种机器，就意味着受到行为规律的制约，这也是正确的。然而，我们不能像认为完备的指导规则不存在一样，简单地认为完备的行为规律也不存在。就我们所知，发现这种规律的唯一方法是科学观察，我们当然也明白，在任何情况下，都无法说："我们的搜寻已经很充分，并没有这样的规律。"

我们可以作出更为有力的证明：任何这种形式的陈述都是未经证实的。让我们假定，如果这种规律存在，我们就有把握找到它。那么，如果有一台离散状态的机器，就完全有可能在

一个合理的时间段内，比如说在一千年里，由对它的充分观察得到的发现，来预见它未来的行为。但是看来事情并非如此。我在曼彻斯特计算机上装了一个小程序，仅用了1000个存储单元，然后，给这台机器一个16位数，它在两秒钟内就用另一个16位数作出答复。我会认为，没有人因熟知这些对程序来说是充分的答复，而能够预言任何对未经试验的值所作的答复。

有关超感知觉的论点

我假定读者对超感知觉的思想是熟悉的，也熟悉它所指的四个方面，即：心灵感应、视力穿透、预知未来和远距离致动。这些令人眼花缭乱的现象，似乎是对我们所有常规科学思想的否定。我多么希望证明这些都是假的！遗憾的是，至少对心灵感应，统计学提供了强有力的证据。让人们改变思想，接受这些新的事实，是十分困难的。人们一旦接受了这些思想，似乎离相信鬼怪也就不远了。作为第一步，我们先来看看这种思想：我们的身体是直接根据已知的物理定律，以及另一些尚未发现、但也多少有些类似的定律运动的。

我认为这是一个很有分量的论证。人们可以回答说，许多科学理论尽管与超感知觉相抵触，在实践中看来仍是行之有效的。还可以说，如果忘记超感知觉，事实上是可以过得很好的。这是一种没有什么作用的安慰，人们所担心的是，思维恰恰是那种与超感知觉有特殊联系的现象。

在超感知觉的基础上，可能出现下面这种较为特别的论点："我们来做模仿游戏，作为参试者的是一个擅长接受心灵感应的人和一台数字计算机。提问者可以问这样的问题：'我

手上的这张牌属于什么花色?'这个人通过心灵感应或视力穿透,在400张牌中作出正确回答130次。机器只能任意猜测,可能有104次正确,于是提问者就能作出正确判断。"这里出现的是一个值得注意的概率,假定数字计算机带有随机数字发生器,自然会用它来决定作出什么回答。但是随机数字发生器也可能受到提问者远距离致动作用力的影响。这种远距离致动有可能使机器猜对的次数多于由概率估算的次数,所以提问者还是不能作出正确判断。另一方面,提问者也有可能根本不作提问,而是通过视力穿透作出正确的猜测。对于超感知觉来说,任何事情都可能发生。

如果承认有心灵感应,我们的试验必须严加控制。可以认为这情况类似于:提问者正在自言自语,一个参赛者把耳朵贴在墙上倾听。只有把参赛者关在"防心灵感应室"内,才能完全符合要求。

7. 学习机

读者可能早就想到了,我无法从肯定的角度作出非常有说服力的论证来支持我的观点。如果能拿出的话,我就不用这样煞费苦心地列举反面观点中的错误了。不过我还是有一点证据,现在就来谈一谈。

此刻我们再回到洛夫莱斯夫人的反对意见上,这个意见认为机器只能做我们叫它去做的事情。有的人说,人能够把一个思想"注入"机器,机器也会在一定程度上作出反应,然后回归静止,就像钢琴弦被小锤敲了一下那样。另一种比喻是

小于临界尺寸的原子反应堆：注入的思想像一个从外面进入反应堆的中子。每一个这样的中子都会引起一定扰动，最后再归于沉寂。然而，反应堆的尺寸增加到一定程度，这个侵入中子引起的扰动就很可能继续进行，不断增大，直到整个反应堆被摧毁。对心灵来说，有类似的现象存在吗？对机器来说呢？对人类心灵来说，确乎有这样的现象。大多数心灵状态看来是“亚临界的”，即相当于这一比喻中所说的低于临界尺寸的堆。一个出现在这种心灵中的思想，所引起的思想响应平均起来还不到一个。超临界的情况所占比例甚小。一个出现在这种心灵中的思想，有可能引发整个“理论”，其中包含着第二层、第三层和更远距离的思想。动物的心灵看来毫无疑问是亚临界的。从这种比喻出发，我们要问：“可以使机器成为超临界的吗？”

“洋葱皮”比喻也有助于我们的理解。在考虑心灵或大脑功能时，我们发现某些运算可以用纯机械方式来解释。这种情况不能代表真正的心灵，它是一种表皮，如果我们要找到真正的心灵，就必须把它剥掉。但是在剩余的部分中，我们又会发现新的要剥去的表皮，这种做法可继续下去。照此进行，我们是抵达了“真正”的心灵呢，还是最终只看到一层内中空空如也的皮？如果是后者，整个心灵就是机械的。（然而它不是一台离散状态机器，这一点我们已讨论过。）

以上两段算不上令人信服的论证，把它们看作是“抛砖引玉”也许更为得当。

能够对§6开始时谈到的那个观点提供真正令人满意的根据，恐怕要等到本世纪末了，那时将由上述实验来说话。但是在这期间，我们能说些什么呢？如果要使这实验成功，我们

现在应采取什么步骤呢?

正如我已经解释过的那样,这主要是一个编程问题。工程方面的改进也是必要的,但看起来还能适应要求。估计大脑的存储容量在 10^{10}到 10^{15}个二进制数字之间。我倾向于较低的值,并且认为用于较高级思维的只是其中的一小部分。它们之中的大部分可能是用于保留视觉印象的。如果成功表演模仿游戏所需的容量超过 10^9,我会感到吃惊,至少与盲人的情况不符。(注:第 11 版《大不例颠百科全书》的容量是 2×10^9。)即使采用目前的技术,也完全有可能实现 10^7 的存储容量。也许根本没有必要提高机器的运算速率。那些可当作神经细胞对应物的现代机器部件,它们的工作大约比神经细胞快一千倍。这样就有了一个“安全裕度”,可以弥补各种情况引起的速率损失。于是我们的问题就是弄清楚该怎样为这些机器编制做这个游戏的程序。根据我目前的工作速度,我大概一天能编一千个数字的程序,所以如果不出现失误,大约 60 个人员,持续工作 50 年,才有可能完成这项工作。看来需要某种更迅捷的方法。

在试图模仿成人心灵的过程中,我们必须对那个把心灵带入它所处状态的过程作大量的思考。我们应当注意三个因素:

1. 心灵的初始状态,比如说出生时的状态;

2. 所受教育;

3. 心灵经历的、但不能看作是教育的其他经验。

与其试图编制模拟成人心灵的程序,为什么不代之以编制模拟儿童心灵的程序呢? 如果让儿童的大脑经历适当的教育过程,就能获得成人的大脑。可以假定儿童的大脑多少有

点像刚从文具店买来的笔记本,它的机制很少,但有大量空白页。(根据我们的观点,机制和书写差不多是同义的。)我们希望的是,儿童大脑的机制非常之少,很容易为类似的东西编制程序。作为初步近似,我们可以假定,对其实施教育的工作量,与教育人类儿童所用的一样多。

这样,我们就把问题分作两部分:儿童心灵程序和教育过程。这两个部分的联系十分密切。我们不能指望在初次尝试中就得到性能良好的儿童机。我们必须对一台像这样的机器进行教学实验,看一看它学习得如何,然后再试另一台机器,看它比前一台是好还是差。这一过程和进化显然存在联系,如下列等式所示:

儿童机的结构 = 遗传素质

儿童机的变化 = 突变

实验者的判断 = 自然选择

然而,人们也许希望这个过程能比进化更迅捷。用适者生存作为衡量优势的方法太缓慢。通过智能练习,实验者有可能使进度加快。实验者不受随机突变的限制,这个事实也同样重要。如果能够追踪某个弱点的原因,实验者就有可能设计出一种突变来改进它。

教机器学习的过程,不可能与教普通儿童的完全相同。例如,没有给机器安上腿,所以不能要求它到外面去把煤斗装满。它也可能没有眼睛。尽管这些缺陷可以由精巧的技术工艺来弥补,但是把这个杰作送到学校去,不会不引起其他孩子对它的大肆取笑。必须对它作某种个别辅导。我们不必过分看重腿、眼睛什么的,海伦·凯勒小姐的例子表明,只要能够用这种或那种方法在老师和学生之间实现双向交流,教育就能

够实现。

我们通常把惩罚和奖励与教育过程结合起来。有些简单的儿童机可以根据这种原理构造或编程。机器必须构造得使那些在出现惩罚信号前不久发生的事件不大可能再次发生，而奖励信号则能提高导致这信号的事件的重复概率。这些规定并没有预先假定机器部件有任何感知方式。我曾在一台这样的儿童机上做过一些实验，成功地教会它一些事情，但是教授方法太不正规了，所以这个实验还不能算是真正成功的。

采用惩罚和奖励充其量只能是教学过程的一部分。粗略地说，如果教师没有其他与学生沟通的方法，那么能教给学生的信息量，不会超过所施奖励和惩罚的总数。在儿童学会重复"Casabianca"这个词时，如果其内容只能通过"二十问"的技巧获知，儿童很可能感到十分痛苦，因为每一个"否"字都是一次打击。所以必须有某种别的"非情感"交流渠道。如果有这种渠道存在，就有可能通过奖惩方式教机器学会服从用某种语言，例如符号语言下达的命令。这些命令是经过"非情感"渠道传递的。使用这种语言，将会大大减少所需的奖惩次数。

至于何种复杂程度适合于儿童机，有各种不同的见解。可以尝试使它尽可能地简单，只要不违背一般原理即可。另一种做法则是，可以用一个完全的"嵌入"① 逻辑推理的系统。在后一情况下，存储主要被定义和命题所占据。命题状况类型各异，例如：完全确认的事实、猜想、由数学证明的定理、权威的意见、具有逻辑命题形式而不具有置信值的表达

① 也可以说是"内编程"，因为我们的儿童机是在数字计算机内编程的。但是，逻辑系统不是必须学习的。

式。某些命题可以看作是“命令”。机器应按这种方式构造：只要一个命令被归入“完全确立”一类，就会有恰当的动作自动发生。为了说明这一点，假定老师对机器说，“现在做你的家庭作业。”这样，就会使“老师说‘现在做你的家庭作业’”包括在“完全确立”的事实中。“老师所说的每件事都是正确的”，也属于这样的事实。把这两者结合起来，最终会导致一个命令：“现在做你的家庭作业”，这个命令也包括在完全确立的事实中，同时，根据机器的构造，这就意味着实际上已经开始做家庭作业，而结果是非常令人满意的。机器所用的推理过程，没有必要得到最严格的逻辑学家的首肯。例如，可能不存在类型的层级体系。但是这也并不是说就会发生类型错误，这有点像我们越过无护栏的悬崖，也不一定会坠落一样。适合的命令（在系统内部表达的，但并不形成系统所具有的规则的一部分）例如“一个类别，如果不是老师已经提到的那个类别的子类，就不要使用”，与“不要走得太靠边”有异曲同工之效。

一台没有肢体的机器所能执行的命令，必然像上述例子（做家庭作业）那样，具有一定的智能特征。在这些命令之中，占重要地位的是那些规定有关逻辑系统规则的实施顺序的命令。因为在人们使用逻辑系统的每一阶段上，都有非常之多的步骤可供选择，就遵循逻辑系统的规则而言，无论应用其中哪一个，都是许可的。这些选择造成了有才华的和无能的推理者之间的差异，但这不是正确的和谬误的推理者之间的差异。能够产生这种命令的命题，可能是“提到苏格拉底时，用三段论第一格的第一式”，或者“如果已经证明了一个方法比另一个方法快，就不要使用那个慢方法”。这些命令中有的可

能是“权威制定的”，而另一些可能是机器自身产生的，例如通过科学归纳产生。

有关学习机的思想，在有些读者看来，可能是自相矛盾的。机器的运算规则怎么可能改变呢？这些规则应该完备地描述机器的反应，无论它的历史可能是什么，无论它会经历什么变化。因此这些规则在时间上是相当恒定的。的确是这样。在学习过程中发生变化的规则，是那种力量不大、只是暂时有效的规则，这就是对于这种自相矛盾的解释。读者在美国宪法中也可以找到类似的东西。

学习机的一个重要特征是，它的教师对于它的内部过程是怎样的，往往知之甚少，尽管他在一定程度上仍能预言他学生的行为。这一特征非常适用于一台其设计(或程序)经过充分试验的儿童机所产生的机器的后续教育。这样做，显然有别于用机器做计算时的正常过程，在后一种情况下，人们的目的是对机器在计算中的每一瞬间的状态，都给以清晰的有关智力的描述。这个目的要经过艰苦的努力才能达到。在这样的事实面前，那种认为“机器只是能做那些我们知道怎样命令它去做的事情”① 的观点，就显得不可思议了。我们能够装入机器的大多数程序，都会使机器做一些我们根本无法意识到的事情，或是我们认为是完全随意的行为的事情。可以假定，智能行为存在于对计算所含照章行事做法的偏离中，不过只是小量的偏离，不会引起随意行为或没有意义的重复循环。通过教和学的过程，从为我们的机器参加模仿游戏做准备中得出的另一个重要结果是，“人类的易错性”很可能是通过一

① 与洛夫莱斯夫人的说法相比，这里增加了“只是”(only)一词。

种相当自然的方式消除的，也就是不用作专门“训练”，就可以消除。（读者应把这与“有关各种缺陷的论点”一节中的观点协调起来。）由学习得到的过程不会产生一个百分之百的必然结果，如果能产生，这些过程倒是不能不学习的。

明智的做法可能是在学习机中设置一个随机单元。在我们搜索某个问题的解时，随机单元是相当有用的。例如，假定我们要在 50 到 200 之间寻找一个数字，这个数字与它数位之和的平方相等，我们可以从 51 开始，然后试 52，并继续下去，直到找到那个有效的数字。换一种方式，我们也可以随机地选出数字，直到选出那个正确的数字。这个方法的优点是，不需要记录已经试过的值，而缺点是同一数字可能试两次，但是在有几个解的时候，这就无所谓了。而系统方法的缺点是，可能不得不先去试探很大一片根本没有解的区域。现在，学习过程可以看作是寻找一种让老师（或某种别的判据）得到满足的行为形式。既然可能有非常之多的合乎要求的解存在，随机方法看来比系统方法更好一些。应该看到，在模拟进化过程中，用的就是这个方法。但是那里不可能用系统方法。人们怎么可能将已试验过的不同基因组合全都记录下，以避免再做重复的试验呢？

我们或许期待着，有一天，机器能够在所有纯智能的领域中同人类竞争。但是从哪里起步最好呢？甚至这也是一个困难的抉择。许多人认为，非常抽象的活动，比如下棋，可能是最好的。也有人认为，最好是给机器配备能买得到的最好的感觉器官，然后教它懂英语，并讲英语。这个过程可以像通常教孩子那样，指着东西，说出它们的名字，等等。我还是不知道正确答案是什么，但是我认为两种方法都应该试一试。

我们的目光所及，只能在不远的前方，但是可以看到，那里有大量需要去做的工作。

参考书目

Church, A. (1936). 'An Unsolvable Problem of Elementary Number Theory.' *American J. Mathematics* 58: 345–63.

Gödel, K. (1931). 'Uber Formal Unentscheidbare Sätze der Principia Mathematica und Verwandter Systeme, I.' *Monatshefte für Mathematica und Physics*, pp. 173–89.

Hartree, D. R. (1949). *Calculating Instruments and Machines*. Urbana: University of Illinois Press.

Kleene, S. C. (1935). 'General Recursive Functions of Natural Numbers.' *American J. Mathematics* 57: 153–7, 219–44.

Russell, B. (1945). *History of Western Philosophy*. New York: Simon and Schuster.

Turing, A. M. (1937). 'On Computable Numbers, with an Application to the *Entscheidungsproblem*.' *Proc. London Math. Soc.* 43: 544; also 42 (1936): 230–65.



心灵、大脑与程序

J·R·塞尔*

我们应当怎样评价近来计算机在模拟人类认知能力方面的成果所具有的心理学和哲学的意义呢？在回答这个问题时，我发现，将我称之为“强”AI的东西与“弱”AI或“审慎的”AI加以区别是有益的。就“弱”AI而言，计算机在心灵研究中的主要价值是为我们提供了一个强有力的工具。例如，它能使我们以更严格、更精确的方式对一些假设进行系统阐述和检验。但是就“强”AI而言，计算机不只是研究心灵的工具，更确切地说，带有正确程序的计算机确实可被认为具有理解和其他认知状态，在这个意义上，恰当编程的计算机其实就是一个心灵。在强AI中，由于编程的计算机具有认知状态，这些程序不仅是我们可用来检验心理解释的工具，而且本身就是一种解释。

我不反对弱AI的论断，至少就本文来说是如此。我这里的讨论针对的是上面定义为强AI的那种论断，特别是这样的论断：恰当编程的计算机确实具有认知状态，因而程序就是对人类认知的解释。我以后提到AI时，都指这个强版本，如上面两种论断表达的那样。

我将考察R·尚克和他的耶鲁同事们的工作（Schank and

Abelson 1977)，因为与其他类似的论断比起来，我对此更为熟悉，同时在我想要考察的这类工作中，它可作为一个很清楚的例子。但是以下内容并不拘泥于尚克程序的细节。这些论证同样适用于威诺格拉德的 SHRDLU(Winograd 1973)，魏曾鲍姆的 ELIZA(Weizenbaum 1965)，当然还有图灵机对人类心理现象的各种模拟。

抛开各种细节，简单地说，尚克的程序可描述如下：程序的目的是模拟人类理解故事的能力。人类理解故事的能力具有这样的特点：在回答有关故事的问题时，即使所给信息从未在故事中直接提到，他们也有能力回答。举例来说，假如你听到这样一个故事："一个人走进一家餐馆，要了一份汉堡包。汉堡包送上来时，已被烤焦，这个人生气地冲出餐馆，没有付账，也没有留下小费。"现在，如果问你："这个人吃汉堡包了吗？"你会根据推测回答："不，他没有吃。"同样，你听了下面的故事："一个人走进一家餐馆，要了一份汉堡包。汉堡包送上来时，他感到很满意，离开餐馆时，在付账之前，他给了女服务员一大笔小费。"再问你这问题："这个人吃汉堡包了吗？"根据推测，你会回答："是的，他吃了汉堡包。"现在，尚克的机器可以用同样的方式对有关餐馆的问题作出类似的回答。为了做到这一点，这些机器具有人类所具有的那种关于餐馆信息的"表述"，它能使机器在听到这类故事后，作出如上的回答。先给机器一个故事，然后让它回答问题，机器打印出来的正是人类听了同样故事之后会作出的那种回答。强 AI 一派的人断

---

* J·R·塞尔，"心灵、大脑与程序"，引自《行为和大脑科学》3(1980)，第 417—424 页，剑桥大学出版社，1980。剑桥大学出版社允许作者重印。

J·R·塞尔(John R. Searle)，加利福尼亚大学(伯克利)哲学教授。

言，在这种一连串的问答中，机器不仅仅是在模拟人类的能力，同时：

1．完全可以说机器**理解**这个故事，并为许多问题提供了答案；并且，

2．机器和它的程序所完成的工作，是对人类理解故事和就故事回答问题的能力的**解释**。

在我看来，尚克的著作[①] 全然未证实这两个论断，下面我来说明这一点。

无论对于什么心灵理论，检验它的方法，就是问一问自己，如果我们的心灵实际按照这一理论所说的那种所有心灵都采用的原则去工作，将会出现什么样的情况。我们就通过下面要讲到的**思想实验**，将这种检验应用于尚克的程序。假定我被锁在一间屋子里，并给了我一大批中文文本；而且，假定我对中文一窍不通（事实也是如此），既不会写，也不会说，甚至我也没有把握，在辨认中文文本时能否把中文文本同日文文本或无意义的曲线区分开来。对我来说，中文文本和许多无意义的曲线简直一模一样。再假定，在第一批中文文本之后，接着又给了我第二批中文脚本，并带有一套规则，使第二批与第一批发生联系。规则是用英文写的，我和其他以英文为母语的人一样是理解这些规则的。用这些规则，我可以把一组形式符号与另一组形式符号联系起来，这里“形式”的意思只是说，我根据这些符号的形状就完全可以确认它们。现在，假定又给了我第三批中文符号，同时还有一些指令，仍是英文的，这些指令使我可以把第三批的元素同前两批联系

---

① 当然，我不是说尚克本人提出了这两个论断。

起来，并指示我怎样送回某种特定形状的中文符号，作为对第三批中送给我的那些特定形状符号的响应。给我所有这些符号的人，我并不认识，他们把第一批符号叫“脚本”，第二批符号叫“故事”，第三批符号叫“问题”，而且把我送回响应第三批文本的符号叫作“对问题的回答”，同时，把他们给我的那套英文规则叫作“程序”。现在，让故事变得稍微复杂一点，设想这些人又给了我一些我所理解的英文故事，然后他们用英文问了我一些关于这些故事的问题，我也用英文回答他们。又假定，经过一段时间，我变得擅长遵循指令来处理中文符号，同时程序员也变得擅长编写程序，以致从外部来看，也就是据我被关屋外的那些人来看，我对问题的回答与讲中文母语的人的回答毫无区别。凡是看过我的回答的人，谁也不会知道我一个中文字也讲不了。让我们再假定，我对英文问题的回答，与其他讲英文母语的人也没有区别，这当然是毫无疑问的，理由很简单，我本来就是一个以英文为母语的人。从外部来看，也就是在那些读了我的“回答”的人看来，我对中文问题和英文问题回答得同样好。但是与英文的情况不同，在中文的场合，我是通过处理不理解的中文符号而得出答案的。对中国人来说，我的行为简直像是一台计算机。我是根据形式上规定好的元素来执行计算操作的。就中国人的目的而言，我不过例示了一个计算机程序。

强 AI 所作的论断是，编程的计算机理解这些故事，同时，这个程序在某种意义上解释了人类的理解。现在我们要做的是，根据我们的思想实验来审视这些论断。

1. 关于第一个论断，我认为，从我丝毫不理解中文故事的例子来看，这一点已经相当清楚。我的输入和输出，与讲中

文母语的人没有区别，而且你想要任何形式的程序，我都可以有，但我仍旧什么也不理解。根据同样的理由，任何故事，无论它们是中文的，英文的，或是其他文字的，尚克的计算机一概都不理解，因为在中文的情况下，计算机就是我，在计算机不是我的情况下，与我什么都不理解的情况相比，它同样地无知。

2. 关于第二个论断，即程序解释了人类的理解，我们不难看到，计算机及其程序并没有提供理解的充分条件，因为计算机和程序正在运行，这中间不存在理解。但是它是否为理解提供了必要条件或重要帮助呢？强 AI 的支持者们作出的论断之一是，在我理解英文故事时，我所做的，恰恰与我处理中文符号时所做的相同，或者说相同的成分较多。在英文场合，我的确是在理解，在中文场合则不然，区别这两者的，只不过是更高形式的符号处理。我尚未证明这个论断是错误的，但是从上面的例子来看，它显然是不能令人信服的。这个论断貌似有理，完全出于这一假定：我们能够构造出一个程序，它的输入输出同讲母语的人完全一样；此外还由于假定讲话者具有某个描述层次，在这一层次上他们也是一个程序的例示。基于这两个假定，我们假定，即使尚克的程序不是对理解的完整叙述，至少也是部分叙述。那么我设想它是一种经验的可能性，但是迄今为止没有任何一点理由能让我们相信这是真的，因为上述例子表明(当然不是证明)，计算机程序与我对故事的理解完全是两码事。在中文场合，我具有人工智能能以程序方式输入给我的每样东西，而我什么也不理解；在英文场合，我理解每样东西，但迄今为止尚无任何理由认为：我的理解与计算机程序，即与在由纯形式说明的元素上进行的

计算操作有什么关系。只要程序是根据在由纯形式定义的元素上进行的计算操作来定义的,这个例子就表明了,这些操作本身同理解没有任何有意义的联系。它们当然不是充分条件,也没有任何一点理由认为它们是必要条件,或者它们对理解作出了重要贡献。请注意,这一争论的要点不能简单地归结为,不同的机器在根据不同的形式原理运作时,可以具有相同的输入和输出,这根本不是争论的实质。相反地,无论你把什么样的纯形式原理放入计算机,它们对理解来说都不是充分的,因为人类能够在毫不理解的情况下遵循这些形式原理。没有任何理由认为:这样的原理在理解中是必要的,甚至是有帮助的,因为毫无理由认为:我在理解英文的时候,是在用什么形式程序进行操作。

那么在英文句子场合,我所具备的,而在中文句子场合,所不具备的,是什么呢?显而易见的回答是:我懂得前者的意思,而对后者的意思丝毫不知。但是这到底是什么?而且不管它是什么,为什么我们不能把它给予一台机器呢?这个问题后面再谈,现在先接着讨论上面的例子。

我曾在一些场合把这个例子讲给几个做人工智能研究的人听,有意思的是,他们对这一问题的正确答案似乎有不同的意见。我得到的各种答复之多令人吃惊,所以接下来我想讨论其中那些最普遍的说法(根据它们的地点来源分门别类)。

但是我首先要阻挡一些对于"理解"的普遍误解,因为在许多这种讨论中,人们看到的是关于"理解"一词的一大套高超的手腕。我的批评指出:存在着许多不同的理解程序;"理解"不是一个简单的二元谓词;甚至存在着许多不相同的理解类型和层次,即使排中律也往往不能直接应用于"X 理解 Y"

这种形式的陈述;在很多情况下,究竟 X 是否理解 Y,是一个需要判断的问题,而不是一个简单的事实;如此等等。对所有这些观点,我要说的是:没有问题,当然是这样。但是它们与所争论的观点毫无关系。在一些情况下,“理解”这个词显然确实是适用的,而在另一些情况下,则显然不适用,这两种情况都是我论证时所需要考虑的。[①] 我理解英文故事;我能在较低的程度上理解法文故事;在更低的程度上理解德文故事;根本不理解中文故事。另一方面,我的汽车和我的加法机,什么也不理解:它们对这种事不在行。我们常常运用比喻和比拟的手法,把“理解”和其他认知属性赋予汽车、加法机和其他人造物,但是这种赋予不能说明什么问题。我们说,“门**知道**什么时候应该打开,因为它带有光电管”,“加法机**知道怎样**(**理解怎样**,并且**能够**)做加减法,但对除法则不然”,“恒温器能**感觉到**温度的变化”。我们这样做的原因很有意思,它肯定与我们把自己的意向性[②] 推广到人造物之中这一事实有关。我们的工具是我们的目的的扩展,所以用比喻的方式将意向性赋予它们,是很自然的事。但是我认为,这些例子在哲学上不起任何作用。自动门因带有光电管而“理解指令”的意义,与我理解英文的意义根本不同。如果认为尚克那台编程计算机理解故事的意义,只是就门能理解那种比喻的意义而言,而不是就我理解英文的意义而言,那么这个问题就没有讨论的

---

① 此外,“理解”既指拥有心理(意向的)状态,也指这些状态的真实性(效力,成功)。鉴于这一讨论的宗旨,我们只考虑拥有状态。

② 根据定义,意向性是某种心理状态的特征,由于这种特征,心理状态指向或是涉及世界中的客体和事物状态。例如,信念、欲望和意图是意向性状态,焦虑和抑郁的无指向形式则不是,进一步的讨论见塞尔(Searle 1976b)。

价值。但是纽厄尔和西蒙(Newell and Simon 1963)写道,他们认为计算机的认知类型,与人类的别无二致。我喜欢这个论断的坦率性,这种论断正是我打算进行考察的。我将论证:从严格的意义上讲,编程计算机所理解的,正是汽车和加法机所理解的,就是说,恰恰什么都不理解。计算机的理解并不只是(像我对德语的理解那样)局部的或不完全的,而是零。

现在来看对我那个例子的种种应答:

1. 系统应答(来自伯克利)

"虽然锁在屋中的个体的人不理解故事是真的,但是事实上他只不过是整个系统的一部分,而这个系统确实是理解故事的。这个人面前有一张很大的明细表,上面写着规则,他有大量的草稿纸和铅笔,可以做演算,他还有多组中文符号'数据库'。这样,就不能把理解看成仅是他个人的事,而应看成是他作为一部分的那一系统整体的事。"

对于这种系统理论,我的回答十分简单:让个体使系统的所有这些元素内化。他要牢记明细表中的规则,和中文符号数据库,并且在自己的头脑中做所有演算。于是个体就收纳了整个系统,把这个系统内的东西全都包罗进来。我们甚至可以摆脱房间,设想他在室外工作。然而一切如故,他还是一点也不理解中文,更不用说,系统也不理解,因为个体中不存在的东西,在系统中仍然不存在。如果个体未作出理解,就不存在系统能够作出理解的方式,因为系统只是个体的一部分。

事实上,甚至对系统理论做这种回答我都觉得有点难堪,

因为在我看来,这个理论从一开始就是不合理的。其思想是,在一个人不理解中文的时候,这个人和一些纸张的结合,或许能以某种方式理解中文。我很难想象,一个人如果不掌握某种思想体系,怎么会觉得这种思想当真可信。可是,我认为许多信奉强 AI 思想体系的人,最终会倒向与之非常相似的说法,所以我想就此进一步作些讨论。这个观点的一种版本是,内化系统例子中的那个人不理解中文,是相对于讲中文母语的人理解中文而言的(例如,就像他不知道这个故事与餐馆和汉堡包有关一样,等等),然而,"这个人作为一个形式符号处理系统",是的确理解中文的。这个子系统属于作为中文形式符号处理系统的人,不应当与英文子系统混淆。

所以在这个人身上确实有两个子系统,一个理解英文,另一个理解中文,同时"这两个系统之间没有什么联系"。但是我要回答说,它们之间不仅几乎没有什么联系,它们简直毫无共同之处。理解英文的子系统(假定我们可以暂时使用"子系统"这个行话来谈论)知道那些故事是关于餐馆和吃汉堡包的,也知道自己正在接受关于餐馆的提问,并且正在根据故事的内容做各种推理,尽力而为地回答问题,等等。但是中文系统对此一无所知。英文系统知道"汉堡包"表示汉堡包,而中文子系统只知道"甲"的后面是"乙"。他知道的仅仅只是,在一端,各种各样的形式符号被引入,根据用英文写成的规则进行处理,然后在另一端,另一些符号被送出。而我的那个例子的全部作用就是要论证:这种符号处理本身,对于任何真正意义上的中文理解来说,都不可能是充分的,因为这个人只会在"甲"后面写出"乙",而非用中文理解任何东西。在人体内假定一些子系统,并不足

以证明这个论点，因为子系统的处境，比起前面那个人来，也强不了多少。它们所具有的东西，仍然与讲英文的人（或子系统）所具有的东西毫无共同之处。的确，在所描述的这一情况中，中文子系统只是英文子系统的一部分，一个根据英文规则进行无意义的符号处理的部分。

首先，我们来问问自己，为了达到系统应答的目的，作出了什么假定，也就是说，在这一应答中，为了说明在当事人的内部必然有一个真正理解中文故事的子系统，要假定哪些**独立**的根据？就我所知，仅有的根据是，在这个例子中，我与讲中文母语的人有完全一样的输入和输出，和一个能从一种语言到另一语言的程序。但是，纵观这些例子，不过是证明了：这种情况不足以提供在我理解英文故事意义上的理解，因为一个人也好，组成一个人的一组系统也好，可以把输入、输出和程序正确地结合起来，但是在与我理解英文相称的真正意义上，依然什么也不理解。说我内部**必然**有一个理解中文的子系统的唯一起因，是我有一个程序，并且我通过了图灵检验，我可以骗过讲中文母语的人。但是争论点之一恰恰是图灵检验的恰当性。这个例子表明，可以有两个“系统”，它们都通过了图灵检验，但是其中只有一个系统理解。如果说，因为两者都通过了图灵检验，所以它们必然都能够理解，这根本构不成反对该观点的论据。因为这一论断不能满足我身上那个理解英文的系统大大超出那个只会处理中文的系统的论据。简单说，系统应答相当武断，在缺乏论据的情况下坚持认为该系统必定理解中文。

进而言之，系统应答看来会导致交相影响的混乱结果。如果我们根据我具备某种类型的输入和输出，以及处于中间

的程序，就下结论说，我身上必然有认知存在，那就好像所有非认知类型的子系统都将变成认知类型的。例如，我的胃在一定描述层次上进行信息加工，它就是不计其数的计算机程序的一个例示，但是我认为我们不应该说它有任何理解力(参阅 Pylyshyn 1980)。但是，如果我们接受系统应答，那么就难以明白，我们怎么能避免说胃、心脏、肝脏等等都是具有理解力的子系统，因为根本就不存在一种原则的方式，可以把说中文系统理解与说胃理解的缘由区别开来。还可以指出，说中文系统以信息作为输入输出，而胃以食物和食物生成物作为输入输出，并不是对这个问题的回答，因为从当事人来看，也就是从我来看，无论在食物里，还是在中文里，都没有信息——中文只不过是许多没有意义的曲线。在中文的场合，信息只存在于程序员和翻译者的眼中。同时也没办法阻止他们把我的消化器官的输入输出看成是信息，如果他们想要这样做的话。

这最后一点与强 AI 中的一些独特的问题有关，因此有必要暂时扯开去把它解释一下。如果强 AI 要成为心理学的一个分支，它必须能够区别真正的心理系统与非心理系统。它也必须能够区别心灵的工作原理与非心灵系统的工作原理，否则它不能向我们解释什么是关于心理的特殊心理成分。心理与非心理的区分不能仅仅存在于观看者的眼中，它必须是系统所固有的。否则会出现这样的情况：一个观看者可以根据他的喜好，把人看作是非心理的，而把例如飓风看作是心理的。但是在 AI 文献中，这种区分常常以某些方式被抹煞了，这种做法最终对认为 AI 是认知探索的主张产生出恶劣的影响。例如，麦卡锡写道："一些机器，即使像恒温器那样简单，

也可以看作是有信念的,而有信念,看来是大多数能够表现出问题求解能力机器的特点。"(McCarthy 1979)每一个认为强 AI 有可能作为一种心灵理论的人,都应当对这个评论的含义进行深思。有人要我们把这种看法作为强 AI 的一个发现:墙上那块用来调节温度的金属板有信念,与我们以及我们的配偶和孩子们有信念,具有完全等同的意义,不仅如此,房间里"大多数"其他机器——电话、录音机、加法机、电灯开关,也在这种真正的意义上具有信念。本文的目的不是与麦卡锡的观点进行争论,所以我不加论证地直接提出如下看法。心灵研究是以人类具有信念,而恒温器、电话和加法机没有信念这些事实作为起点的。如果你提出一个理论否认这一点,你已经给这理论提供了一个反例,从而这理论就是错误的。我们的印象是,在 AI 中,有这类观点的人,以为他们能混过去,因为他们并没有真正严肃地思考过它,同时他们认为别的人也不会严肃地思考它。我提议,至少用片刻时间,对它严肃思考一下。用一点时间努力想一想,必须用什么来证实那边墙上的那块金属板具有真正的信念,这种信念有适从方向、命题内容和满足条件;这些信念可能是强信念也可能是弱信念;紧张、焦虑或安全的信念;武断、理性或迷信的信念;盲目信任或优柔寡断;以及任何一种信念。恒温器不是候选者,胃、肝脏、加法机或电话也不是候选者。但是,既然我们严肃地对待这一思想,就应当注意到,要使强 AI 论断成为一门心灵科学,这一思想的真实性是至关重要的,因为现在到处都是心灵。我们要知道的是,使心灵与恒温器和肝脏区别开来的是什么。即使麦卡锡是正确的,也不要指望强 AI 会告诉我们这一点。

2．机器人应答(来自耶鲁)

“设想我们编写了一个与尚克程序类型不同的程序，同时，我把一台计算机放入机器人中，这台计算机不只是接收形式符号作为输入，发送形式符号作为输出，它实际上是在操作机器人，通过操作，使机器人做出与感知、步行、走来走去、钉钉子、吃东西、饮水十分相像的事情——以及你希望的任何事情。例如，机器人配有一台摄像机，它就能‘看见’，有了手臂和腿，它就能‘动作’，而这一切都由计算机‘大脑’来控制。这种机器人与尚克的计算机不同，具有真正的理解力和其他心理状态。”

关于机器人应答，首先应注意的是：由于这个应答增加了一套与外部世界的因果联系，它不言而喻地承认了，认知并不仅仅是一个形式符号处理的问题(参阅 Fodor 1980)。但是，我对机器人应答的回答是：增加了这种“感知”和“运动”能力，没有特别地在理解方面或者一般地在意向性方面给尚克原来的程序增加什么东西。为了看清这一点，应注意到同样的思想实验也适用于机器人的情况。假定你不是把计算机放进机器人中，而是把我放进房间里，像原先读中文的例子那样，你给了我更多的中文符号，以及更多的英文指令，让我把这些中文符号彼此匹配起来，并向外反馈中文符号。假定在我不知道的情况下，给我的某些中文符号来自机器人所带的摄像机，而我送出的另一些中文符号能使机器人中的发动机驱动机器人的腿或手臂。强调这一点是重要的：我所做的一切都是处理

形式符号——除此而外，我对其他事实一无所知。我从机器人的“感知”装置接收“信息”，向它的动力装置发送“指令”，而对这两种事实都不知情。我是机器人中的小精灵，但与传说中的小精灵不同，我不知道正在发生什么事。除了符号处理规则之外，我什么也不理解。我想通过这个例子说明：机器人根本没有意向状态，它只是受电路和程序支配简单地来回运动而已。而且，如程序例示的那样，我也没有任何相关类型的意向状态。我做的一切，不过是遵循有关处理形式符号的指令而已。

3. 大脑模拟者应答
（来自伯克利和麻省理工学院）

“假定我们设计一个程序，它并不表述我们具有的外部世界的信息，如尚克脚本中的那些信息，但是它模拟一个讲中文母语的人在理解中文故事和对故事作出回答时在他大脑突触上的神经元激发的实际顺序。这台机器接收中文故事和有关故事的问题作为输入，模拟实际懂中文的大脑在加工这些故事时所具有的形式结构，并以中文回答作为输出。我们甚至可以设想，这台机器运作时，用的不是单一串行程序，而是并行运作的整组程序，也就是我们认为实际人类大脑在加工自然语言时采用的那种操作方式。毫无疑问，在这种情况下，我们不得不说，机器理解这些故事。如果我们拒绝承认这一点，我们是不是也要否认讲中文母语的人理解这些故事呢？在突触层次上，计算机程序和懂中文大脑的程序会有、或

者能有什么差别呢?”

在反驳这一应答之前,我想扯开去指出,任何人工智能(或功能主义等)派别作出这样的应答是不近情理的,因为我认为强 AI 的全部思想就是:要了解心灵是怎样工作的,并不需要了解大脑是怎样工作的。这一基本假设是,或据我的猜想是:存在着一个心理运作层次,它是由形式元素上的计算过程组成的,这些过程构成了心灵的本质,并能通过所有不同类型的大脑过程来实现,正如任何计算机程序都可以通过不同的计算机硬件来实现一样。根据强 AI 的假定,心灵之于大脑,犹如程序之于硬件,这样,我们就可以不经神经心理学来理解心灵了。如果为了实现 AI,必须知道大脑是怎样工作的,那么我们就不必为 AI 而烦心了。然而,即使我们使 AI 接近于大脑的运作方式,仍然不足以产生理解。为了看清这一点,我们设想让那个只懂一种语言的人去操作一套复杂的带有连接阀门的水管,而不是在房间里摆弄符号。这个人接收到中文符号时,就在用英文写成的程序内查找,看看应当打开和关闭哪些个阀门。每一段水流的连接都对应于懂中文的大脑中的一个突触,整个系统装配起来,使得在全部应有的激发产生后,也就是在所有正确的开关都打开后,中文答案就会在管道系统的输出端冒出来。

那么,这个系统的理解在哪儿呢?它以中文作为输入,它模拟懂中文大脑神经突触的形式结构,并且它给出中文作为输出。但是毫无疑问,这个人是不理解中文的,水管系统也一样。同时,如果我们打算采纳我认为荒谬的观点,即以为人和水管的结合在作某种理解的观点,那么别忘记,从原理上讲,人能够内化水管的形式结构,并在他的想象中完成所有的“神

经元激发”。与大脑模拟者有关的问题是，它模拟的不是大脑的正确东西。只要它模拟的仅仅是突触上神经元激发序列的形式结构，它就没有模拟到大脑的要害问题，也就是大脑的因果特性，大脑产生意向状态的能力。形式特性不能充分说明因果特性，这一点已通过水管的例子得到证明，因为我们可以把全部形式特性与相关的神经生物的因果特性分离开来。

4. 联合应答(来自伯克利和斯坦福)

“虽然前述三个应答中的每一应答单独作为中文屋的反例进行反驳可能没有绝对的说服力，但是如果把这三者结合起来，就会共同形成大得多的说服力，甚至起到决定性作用。我们设想一个机器人，它的头盖骨腔中放着一台大脑形状的计算机，再设想，这台计算机的程序是根据人类大脑的全部突触编成的，同时机器人的整个行为也无法与人的行为相区别，现在把所有这些看作一个统一的系统，而不仅仅是带有输入输出的计算机。在这种情况下，我们当然不得不认为这个系统具有意向性。”

我完全同意，如果我们对此没有更多了解的话，在这种情况下，我们会认为接受机器人有意向性的假设是合理的，也是难以抗拒的。的确，除了外表和行为，结合体的其他部分实在是无足轻重的。如果我们能制造出一个机器人，它的行为与人的行为在一个很大的范围内都没有区别，我们就可以认为它具有意向性，而不去考虑某个相反意见的理由。我们并不需要事先知道，它的计算机大脑是人类大脑形式上的模拟。

但是，我实在看不出这对于强 AI 论断有什么帮助。其原因是，根据强 AI 的观点，例示带有正确输入和输出的形式程序，就是意向性的充分条件——实际是构成意向性的充分条件。正如纽厄尔(Newell 1979)所指出的，心理的本质是对物理符号系统的操作。但是在这个例子中，我们为机器人建立的意向性属性，与形式程序是毫无关系的。它们只是基于这样的假定：如果机器人的举止行为和我们人类充分相似，那么在没有其他相反证据的情况下，我们就认为它必然有像我们一样的心理状态，这种心理状态引起行为发生，又通过行为表现出来，所以机器人必然有一个能够产生这种心理状态的内部机制。如果我们另外知道，不作这种假定时怎样解释它的行为，特别是如果我们知道它有一个形式程序，我们就不会将意向性赋予它了。而这正是我在前面答复第二种反对意见时的观点。

假定我们知道，机器人的行为完全取决于这一事实：在它之中，有一个人正在从它的感觉接收器接收未经解释的形式符号，并把未经解释的形式符号送到它的动力装置，这个人是根据一整套规则来作这种符号处理的。进而再假定，这个人一点也不了解有关机器人的这些事实，他只知道，根据何种无意义的符号，进行何种操作。在这种情况下，我们可以把机器人看成一个精巧的机器傀儡。由此作出这个傀儡有心灵的假设是缺乏真凭实据的，也是没有必要的，因为现在再没有任何理由把意向性赋予机器人，或是赋予它所属的那个系统(当然这不包括人在处理符号时的意向性)。形式符号处理在继续，输入和输出得以正确匹配，但是意向性仅有的真正居所是人，而人对相关的意向状态毫不知情，例如，他看不到进入机器人

眼睛的是什么,他也并不打算使机器人的手臂运动,他也不理解对机器人讲了什么,或者机器人讲了什么。由于前述种种原因,由人和机器人组成的系统,也是什么都不知道。

为了弄清这个观点,可以把这个情况同另一些情况作对比,在那些情况下,我们觉得,把意向性赋予某些别的灵长类动物,像猿和猴子,以及家畜,像狗,是很自然的事。我们所以觉得这是自然的,粗略地说,有两点理由:如果不把意向性赋予这些动物,我们就不能理解它们的行为,同时,我们都知道禽兽是由和我们类似的材料构成的——这是眼睛,这是鼻子,那是皮肤,等等。已知动物的行为具有协调一致性,并假定它是以同样的因果特性材料为基础的,我们就可以作两点假定:动物必然有作为其行为基础的心理状态,同时,产生该心理状态的机制必然是用与我们的材料一样的材料构成的。虽然我们在没有相反理由的情况下确实可以对机器人作类似的假定,但是一旦我们知道这个行为是形式程序的结果,而与物理实体的实际因果特性无关,我们就会放弃意向性的假定(见多个作者 1978)。

对于我的例子还有另外两种回答,它们经常被提到(所以有必要讨论),然而它们实际上是误解了我的观点。

5. 他人心灵应答(来自耶鲁)

“你怎么知道其他人理解中文,或是别的什么呢?仅仅根据他们的行为。现在,计算机能够(原则上)像那些人一样地通过行为测试,所以如果你要把认知属性赋予那些人,

那么原则上，你也必须把认知属性赋予计算机。”

这种反对意见实在不值得花费多少笔墨去回答。这里讨论的问题，并不是我们如何知道别人有认知状态，而是我把认知状态赋予他们时，我是在把什么样的属性赋予了他们。该论证的要点是，认知不可能只是计算过程及其输出，因为在没有认知状态的情况下，计算过程及其输出也可以存在。对这一论证采取视而不见的态度，并不是对它的回答。在“认知科学”中，人们预先假定了心理的实在性和可知性，正像在物理科学中人们不得不预先假定物理对象的实在性和可知性一样。

6. 多重套间应答(来自伯克利)

“你的整个论点预先假定 AI 仅仅与模拟计算机和数字计算机有关。但这恰巧只是当前的技术状况。无论你认为作为意向性本质方面(假定你是正确的)的这些因果过程是什么，我们最终都能制造出具备这些因果过程的装置，而这就是人工智能。所以从你的论点决不会推出人工智能有产生和解释认知的能力。”

对于这一应答，我的确无可反驳，我只能说，由于把强 AI 重新定义为任何人工产生和解释认知的过程，它原先的目标实际上已变得无关紧要。原先为人工智能所作出的那个论断之所以有价值，是因为它是一个精确的、定义完善的论题：心理过程是以形式定义的元素上的计算过程。我曾经关注过对这一论题的质疑。如果该论断重新定义，它就不再是这样的

论题,我的反驳也不再适用,因为适用我反驳的可以检验的假设已经不存在了。

现在,我们返回去看看那个我曾经允诺要设法回答的问题:在我原来那个例子中,假定我理解英文,而不理解中文,从而假定机器既不理解英文,也不理解中文。然而,我之所以理解英文,必然是由我身上的某种因素造成的,同时,我不理解中文,也是因为我身上缺少某种相应的因素,那么,不管这些因素是什么,为什么我们不能把它们给予机器呢?

我明白,从原理上讲,并没有理由说我们不能把理解英文或理解中文的能力赋予机器,因为从某个重要的意义上说,我们带有大脑的身体恰恰就是这样的机器。但是,我也完全清楚,有强有力的证据表明,我们不可能把这种东西给予机器,因为机器的运作完全是根据以形式定义的元素的计算过程确定的。也就是说,机器的运作被定义为计算机程序的例示。我之所以能理解英文,并且具有其他形式的意向性,并不是因为我例示了一个计算机程序(我想,我是不计其数的计算机程序的例示),而就我们所知,这是因为我是某种有机体,具有某种生物结构(即化学和物理结构),在一定条件下,这个结构能够以因果的方式产生感知、行动、理解、学习以及其他意向性现象。本文论点的要点之一就是,只有具有这些因果能力的东西,才可能具有意向性。也许别的物理和化学过程能够产生出完全一样的效果,比如说,火星人也可能有意向性,但是他们的脑是用不同的材料构成的。这是一个经验主义的问题,与光合作用是否能够由化学性质不同于叶绿素的某种物质完成的问题,有其相似之处。

然而,本文论点的中心思想是,从没有一种纯形式模型,

足以凭借其自身产生意向性,因为形式特性自身不能构成意向性,同时它们自身也没有因果能力,它们的能力不过是在例示过程中随着机器运行而产生下一步的形式体系。形式模型的特殊实现方式所具有的任何其他因果特性,与形式模型无关,因为我们总能把同样的形式模型放入显然不存在那些因果特性的别的实现方式中。纵然出现奇迹,讲中文的人精确地实现了尚克的程序,我们也可以把完全相同的程序交给讲英文的人、水管或计算机,尽管它们有程序,它们之中却没有一个理解中文。

在大脑操作中起作用的,并不是突触序列投下的形式影子,而是这些序列的实际特性。我所见到的所有关于人工智能强版本的论点,都主张围绕认知投下的影子画出轮廓,然后断言这些影子是真实的东西。

我想以总结的方式来阐明包含在这个论点之中的某些一般的哲学观点。为清楚起见,我试以问答的方式来进行,并以这个陈旧的问题作为开始:

"机器能够思维吗?"

很显然,回答是"能够"。我们正是这样的机器。

"好的。但是一个人工制品、一台人造的机器能够思维吗?"

假定有可能由人工造出一台带有神经系统的机器,神经元上有轴突和树突,以及其余所有部分,与我们的充分相像,那么对这个问题的回答看来显然仍旧是"能够"。如果你能够精确地复制原因,你就能够复制结果。而且的确有可能使用某种其他种类的化学原理,而不是人类所使用的那些化学原理,产生出意识、意向性,以及其余所有东西。正如我说过的,

这是一个经验主义的问题。

“很好。但是一台数字计算机能够思维吗?”

如果所谓“数字计算机”不是指别的,就是指任何具有某种描述层次的东西,在这个层次上,可以正确地把它描述为例示了一个计算机程序,那么回答当然再一次是能够,因为我们就是不计其数的计算机程序的例示,而我们是能够思维的。

“但是,某个东西**仅仅**因为它是一台带有那类程序的计算机,就能够思维、理解,和做诸如此类的事情吗? 例示一个程序,当然是正确的程序,这本身能够作为理解的充分条件吗?”

我认为,这个问题提得很恰当,虽然它常常同前面的一个或几个问题相混淆。对于这个问题的回答是不能。

“为什么不能?”

因为它们本身所做的形式符号处理没有任何意向性;它们是全然无意义的;它们甚至不是**符号**处理,因为这些符号什么也不代表。用语言学的行话来说,它们只有句法,而没有语义。那种看来似乎是计算机所具有的意向性,只不过存在于为计算机编程和使用计算机的那些人心里,和那些送进输入和解释输出的人的心里。

中文屋例子的目的就是试图说明这一点,它说明只要我们把某些东西放入那个真正具有意向性的系统(一个人)中,同时给他配以形式程序,就能看清形式程序并没有携带任何新增的意向性。例如,它没有给一个人理解中文的能力增加任何东西。

确切地说,AI那个看来如此引人的特征——区分程序与实现方式,对于断言模拟可以作为复制来说,原来是至关重要的。区分程序与它在硬件中的实现方式,似乎可以与区分心

理操作层次与大脑操作层次相提并论。如果我们可以把心理操作层次看作形式程序，那么我们似乎不必通过内省心理学或大脑神经生理学，就可以说明什么是心灵的本质。但是"心灵之于大脑，犹如程序之于硬件"这个等式，在好几点上不能成立，其中的三点如下：

第一，区分程序与实现方式会造成这种结果：同一程序可以用不具有任何意向性形式的各种稀奇古怪的方式来实现。例如，魏曾鲍姆（Weizenbaum 1976: ch.2）详细地说明了怎样用一卷卫生纸和一堆小石子来构造一台计算机。类似地，可以把中文故事理解程序编入一套水管、一组风机或一个只会讲英文的人中，这些事物中没有一个因之而获得对中文的理解。首先，对具备意向性来说，石头、卫生纸、风和水管的材料种类不对头——只有具备与大脑完全一样的因果能力的东西才能够具备意向性，而讲英文的人虽然具有这种对意向性来说是恰当的材料，但不难看出，他并没有因牢记这个程序而获得任何额外的意向性，因为牢记程序这件事不会教他学中文。

第二，程序是纯形式的，但是就这种意义而言意向状态不是形式的。意向状态是根据它们的内容，而不是根据它们的形式定义的。例如，正在下雨的信念不是定义为某种形式外形，而是定义为满足条件的某种心理内容、适从方向（参见Searle 1979a），等等。的确，像这样的信念在句法意义上甚至连形式外形也没有，因为同一个信念在不同语言系统中可能有个数不定的不同句法表达式。

第三，正像我前面提到的，心理状态和心理事件确实是大脑运作的产物，但是程序不是同样方式下的计算机的产物。

"好吧，如果程序全然不构成心理过程，为什么有那么多

的人持有相反的看法呢？这至少需要作出某种解释吧。”

对此，我真不知道该如何回答。首先，认为计算机模拟可能是真事的思想本该值得怀疑，因为计算机其实并不限于模拟心理操作。没有人认为紧急火警的计算机模拟会烧毁邻居的家，或者暴风雨的计算机模拟会把我们大家淋得湿透。可是为什么竟有人会认为对理解的计算机模拟就是真正地理解某个事物呢？有时听到这种说法：要让计算机感觉疼痛，或是堕入情网，是极其困难的事情，但是恋爱和疼痛，既不比认知或什么别的事情困难，也不比它们容易。对于模拟来说，你需要的只是正确的输入和输出，以及处在中间把前者转化为后者的程序。这就是计算机在做任何事情时所具有的一切。把模拟和复制混为一谈，不管是疼痛，恋爱，认知，火警，还是暴风雨，都是犯了同样的错误。

此外，AI 的确曾经看起来（也许不少人现在仍然是这样看）以某种方式再现了心理现象，因而也就解释了心理现象。造成这些错误看法的原因有好几个，我相信，在充分探讨这些原因之前，我们是很难消除这些看法的。

第一，也许是最重要的，是关于“信息加工”这一概念的混乱：在认知科学中，很多人相信，人类大脑及其心灵在做着某种叫作“信息加工”的事情，它可以类比于计算机及其程序所做的信息加工；但另一方面，火警和暴风雨却根本不做信息加工。因此，虽然无论何种过程的形式特征，计算机都能模拟，但是它与心灵和大脑处在一种特殊关系之中，因为在计算机正确编程之后，就理想地带有和大脑一样的程序，在这两种情况下，信息加工是完全等同的，而这样的信息加工确实就是心理的本质。但是这个论点的问题出在“信息”概念的歧义上。

如果“信息加工”的意思是，当人们比如在思考算术题时，或者在阅读故事并回答有关它的提问时，所做的事是“信息加工”，那么，编程计算机所做的事就不是“信息加工”，而是处理形式符号。程序编制者和计算机输出解释者使用符号来替代现实中的物体，这个事实完全是计算机范围之外的事。我要重复一遍，计算机只有句法，而没有语义。因此，如果你在计算机上敲入“2 加 2 等于几?”它就打出“4”，但它根本不知道“4”意味着 4，或是意味着其他任何东西。关键问题并不在于它缺少关于解释它的一阶符号的某种二阶信息，而是对计算机来说，它的一阶符号也没有任何解释。计算机具有的只是更多的符号。因此，引入“信息加工”概念的结果，产生了一种二难推理:或者我们把“信息加工”的概念解释为是指意向性参与了这一过程，或者不这样认为。如果是前者，那么编程计算机就没有做信息加工，它只是在处理形式符号。如果是后者，那么虽然计算机做了信息加工，但是它不过是在与加法机、打字机、胃、恒温器、暴风雨以及飓风做信息加工相同的意义上这样做的。也就是说，它们具有这样一种描述层次，在这层次上，我们可以把它们描述为，在一端接收信息，对信息进行转换，并产生作为输出的信息。但是在这种情况下，外部观察者就必须在通常的意义上把输入和输出解释成信息。无论根据何种信息加工的相似性，在计算机和大脑之间都不可能建立起相似性关系。

第二，在许多 AI 中都存有残余的行为主义和操作主义。由于恰当编程的计算机能具备与人类相似的输入输出模式，我们就试图设定在计算机中有与人类相似的心理状态。但是，我们一旦看到，在某个根本不存在意向性的领域中，一个

系统也可以在概念上和经验上具有像人类一样的能力,我们就会克服这种冲动。我的台式加法机具有计算能力,但是并没有意向性。我在本文中也努力证明了,一个系统可能复制讲中文母语的人所具有的那种输入输出能力,但是仍然不理解中文,不管程序是怎样编制的。图灵检验代表了那种赤裸裸的行为主义和操作主义的传统,而我相信,如果 AI 研究者们完全与行为主义和操作主义断绝关系,存在于模拟与复制之间的许多混淆就会消失。

第三,这种残余的操作主义是同残余的二元论形式联在一起的;的确,强 AI 之所以成立,是由于这样的二元论假定:在与心灵有关的地方,大脑是无关紧要的。在强 AI 中(同样也在功能主义中),重要的东西是程序,而程序又独立于它们在机器中的实现方式。的确,就 AI 而言,同一程序可以由电子机械实现,也可以由笛卡尔的心理实体或是黑格尔的宇宙精神来实现。我认为实际的人类心理现象很可能取决于实际人类大脑的实际物理化学特性,而我的这个思想令很多 AI 研究者震惊不已——这是我在讨论这些问题时得出的一个最出乎意料的发现。但是,如果对此稍加考虑,就会看到,我本不该有出乎意料之感,因为如果不接受某种形式的二元论,强 AI 的计划就无指望实现。这个计划是要通过设计程序来再现和解释心理。但是,只有当心灵不仅在概念上、并且在经验上独立于大脑时,才能实现这个计划,因为程序是完全独立于任何实现方式的。只有相信心灵既在概念上、也在经验上可以同大脑相分离——二元论的强形式,才能希望用编写和运行程序的方式再现心理,因为程序必须独立于大脑,或是任何其他特殊形式的实现方式。如果心理操作在于用形式符号进

行计算操作，那么其结论则是这些操作与大脑没有任何令人感兴趣的联系。唯一的联系是，大脑恰巧是数目不定的许多能例示这一程序的机器类型之一。这种二元论形式并不像传统笛卡尔派断言的那样存在着两类**实体**，但是它在这种意义上是笛卡尔式的：它坚持认为，有关心灵的特殊心理内容与大脑的实际特性没有固有联系。由于 AI 文献中常常包含着对“二元论”的强烈谴责，使我们难以看清这种深层二元论的真正面目；看来那些作者们并不知道，他们的立场预设了一个二元论的强版本。

“机器能够思维吗？”我个人的观点是，只有一种机器能够思维，实际上只有一些类型非常特殊的机器，即大脑和那些与大脑具有相同因果能力的机器，能够思维。而这正是强 AI 在思维问题上几乎没有告诉我们任何东西的主要原因，因为关于机器它没有什么可告诉我们的。根据强 AI 自己的定义，它是关于程序的，而程序不是机器。无论意向性是别的什么东西，它都是一种生物现象。同时，它很可能像泌乳、光合作用或任何其他生物现象一样，与生成它的特定生物化学特性具有因果相关性。谁也不会认为，我们可以通过对泌乳和光合作用中的形式序列进行计算机模拟而得到牛奶和糖类。但是在与心灵有关的地方，由于根深蒂固的二元论，很多人都宁愿相信这样的神话：他们所认为的心灵，其实质是形式加工，它不像牛奶、糖类那样不能独立于非常专门的物质因果关系，它是保持独立的。

为了捍卫这个二元论，人们常常表示出这样的希望：大脑是一台数字计算机（顺便指出，早期的计算机常被称为“电脑”）。但这无济于事。当然，大脑是一台数字计算机，既然每

一样东西都是数字计算机，大脑也不例外。问题是，大脑产生意向性的那种因果能力，并不存在于它例示计算机程序的过程中，因为无论你想要什么程序，都能够由某种东西来例示这个程序，而它并不具有任何心理状态。无论大脑在产生意向性时所做的是什么，都不可能存在于例示程序的过程中，因为没有一个程序凭借自身而对于意向性来说是充分的。①

参考书目

Fodor, J. A. (1980). 'Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology.' *Behavioral and Brain Sciences* 3: 63–110.

McCarthy, J. (1979). 'Ascribing Mental Qualities to Machines.' In M. Ringle (ed.), *Philosophical Perspectives in Artificial Intelligence*, pp. 161–95. Atlantic Highlands, NJ: Humanities Press.

[Multiple authors] (1978). 'Cognition and Consciousness in Non-Human Species.' *Behavioral and Brain Sciences* 1(4): entire issue.

Newell, A. (1979). 'Physical Symbol Systems.' Lecture at the La Jolla Conference on Cognitive Science. Later published in *Cognitive Science* 4 (1980): 135–83.

—— and Simon, H. A. (1963). 'GPS—A Program that Simulates Human Thought.' In E. A. Feigenbaum and J. A. Feldman (eds.), *Computers and Thought*, pp. 279–96. New York: McGraw-Hill.

Pylyshyn, Z. W. (1980). 'Computation and Cognition: Issues in the Foundation of Cognitive Science.' *Behavioral and Brain Sciences* 3: 111–32.

Schank, R. C. and Abelson, R. P: (1977). *Scripts, Plans, Goals, and Understanding*. Hillsdale, NJ: Erlbaum.

Searle, J. R. (1979*a*). 'Intentionality and the Use of Language.' In A. Margolit (ed.), *Meaning and Use*. Dordrecht: Reidel.

—— (1979*b*). 'What is an Intentional State?' *Mind* 88: 74–92.

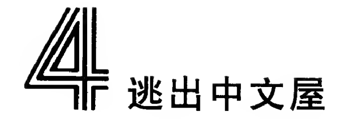
Weizenbaum, J. (1965). 'ELIZA—A Computer Program for the Study of Natural Language Communication Between Man and Machine.' *Commun. ACM* 9: 36–45.

—— (1976). *Computer Power and Human Reason*. San Francisco: W. H. Freeman.

---

① 我同很多的人就这些问题展开过讨论，他们耐心地试图克服我对人工智能的无知，为此我深表感谢。我特别要感谢 N·布洛克、H·德雷福斯、J·豪格兰、R·尚克、R·威伦斯基和 T·威诺格拉德。

Winograd, T. (1973). 'A Procedural Model of Language Understanding.' In R. C. Schank and K. M. Colby (eds.), *Computer Models of Thought and Language*, pp. 152–86. San Francisco: W. H. Freeman.



逃出中文屋

M·A·博登*

J·塞尔在他的文章“心灵、大脑与程序”(Searle 1980)中论证道:心理学中的计算理论基本上是没有价值的。他作出的两个主要论断是:计算理论因其本质上是纯形式的,所以根本不可能有助于我们理解心理过程;同时,计算机硬件也不同于神经蛋白,显然缺乏生成心理过程所需的恰当的因果能力。我要阐明的观点是:这两个论断都是错误的。

他的第一个论断认为如下的流行(形式主义)假定是当然成立的:计算机科学中研究的“计算”是纯句法的,可以(用同样适合于符号逻辑的术语)把它们定义为**应用形式规则,对抽象符号进行的形式处理**。他又说,因此形式主义的论述虽然在解释计算机中无意义的“信息”加工或“符号”处理方面是恰当的,但是不能解释人类心灵是怎样运用**信息**的,或正确地说,是怎样运用所谓**符号**的。意义或是意向性是不能用计算术语来解释的。

这里,塞尔的观点并不是认为任何机器都不能够思维。人类能够思维,而人类也是机器——他接受这一说法。他甚至采用唯物主义的信条,认为只有机器才能够思维。他也没

有说人类和程序是完全不可通约的。他假定,在某一高度抽象的描述层次上,人和任何其他东西一样,都是数字计算机的例示。应该说,他的观点是:任何东西不可能仅仅凭借它例示出计算机程序而思维、表意或理解。

为了让我们接受这一观点,塞尔运用了一个别出心裁的思想实验。他想象自己被锁进一间屋子,屋里有各种各样的纸片,上面有一些涂划。通过窗口,人们可以再递给他有涂划的纸片,他也可以通过窗口把纸片递出。他从一本(英文的)规则书中得知,怎样为这些涂划配对,而这些涂划总是以它们的形状或形式来确认的。塞尔在屋里时,就把他的时间消磨在根据规则去处理这些涂划上。

例如,一条规则指示他:给他递进"甲"时,他应当送出"乙"。规则书中还规定了许多更为复杂的涂划配对序列,但是只有在第一步和最后一步上才涉及到把纸片传递入和传送出房间。在见到某个直接指示他递出一个纸片的规则之前,他也许得找到涂划"丙",并把它同涂划"丁"加以比较,在这种情况下,正是这一比较的结果决定着他所递出的涂划的性质。有时,在见到一个规定他递出什么东西的规则之前,他必须在屋中进行多次这样的涂划与涂划的比较,并完成最后的涂划选择。

就屋中的塞尔而言,"甲"和"乙"只是一些无意义的涂划。可是他不知道,它们都是中文字符,屋外是中国人,他们能对涂划作出中文解释。此外,据中国人的理解,从窗口递进递出

---

* M·A·博登,"逃出中文屋",摘自《心灵的计算机模型》第8章(1988)。剑桥大学出版社允许重印。

M·A·博登(Margaret A. Boden),苏塞克斯大学认知与计算科学学院哲学与心理学教授。

的模式分别是**问题和答案**，因为规则恰好使得大多数问题直接或间接地与他们认为是合理的回答相配对。但是(屋中的)塞尔本人对此一无所知。

根据塞尔的意思，“屋中的塞尔”显然可作为计算机程序的例示，即：他正在完成对未经解释模式的纯形式处理：他完全是句法的，而非语义的。

涂划配对规则等价于**如果-那么**规则，或是“产生式”规则，(例如)在专家系统中通常就是这样使用的。某些内部涂划比较，可以等价于AI研究者在自然语言加工中所说的脚本，例如R·C·尚克和R·P·阿贝尔森(Schank and Abelson 1977)描述的餐馆脚本。在该例中，屋中的塞尔传递纸片的行为，本质上可与尚克的“问题回答”文本分析程序相比较。但是“问题回答”并非是问题回答。屋中的塞尔并非真的在**回答**问题：既然他无法理解这些问题，怎么可能回答呢？实习也无济于事(除了可能提高做涂划配对的速度)，因为屋中的塞尔一旦逃出，与他先前被锁进去时一样，对中文仍一无所知。

当然，屋外的中国人可能认为，供给屋中的塞尔食物和水是有用的，正像在现实生活中，我们愿意花大笔资金建立计算机“咨询”系统一样。但是下面的事实却又是另一回事：已经具备理解力的人，会使用一个原本无意义的形式主义计算系统，来提供那些被他们解释(原文如此)为问题、答案、指称、解释或符号的东西。只有当他们能够从外部说明该形式体系和他们感兴趣的事物之间的映射关系时，他们才会这样做。从原理上讲，同一个形式体系可以映射到几个不同的领域中去，所以(人们)就可以用它回答有关其中任一领域的问题。然而就其本身而言，它可能是无意义的，就像从屋中的塞尔的观点

来看,中文符号是无意义的一样。

塞尔论证道,因此没有一个系统仅仅凭借它例示了计算机程序就能够对某个事物进行理解。因为假如这是可能的,那么屋中的塞尔就能理解中文了。因此,理论心理学以计算概念为基础不可能是正确的。

塞尔的第二个论断关系到对理解的正确解释是什么样的。根据塞尔的看法,应该承认有意义的符号必须由某种具有对生成理解力或意向性来说是"恰当的因果能力"的东西来实现。他说大脑显然具有这种因果能力,而计算机不具有。由于可以在计算机中模拟大脑的组织形式,所以更确切地说,是神经蛋白具有这种能力,而金属和硅不具有:大脑物质的生物化学特性成为关键所在。

广泛引用的 A·纽厄尔关于"物理符号系统"的定义(Newell 1980),遭到塞尔的反对,因为按照这个定义,某种材料只要能够完成形式主义计算,就能够以它来实现符号,而这些计算,公认是能由计算机完成的。在塞尔看来,任何电子计算机都不可能真的处理符号,也不可能真的指称或解释任何东西——**不管**是什么因果相关性将它的内在物理模式同它的行为连接起来的。(这个关于意向性的强实在论观点与 D·C·丹尼特(Dennett 1971)的工具主义恰成对照。根据丹尼特的说法,一个意向性系统是这样的:只要赋予它信念、目标和推理能力,我们就可以解释、预见和控制它的行为。根据这个判据,某些**现存的**计算机程序就是意向性系统,更不用说科幻小说构想出来的那些通人性的机器人了。)

塞尔声称意向性是一个生物学现象。因此它就像光合作用和泌乳一样,只取决于作为其基础的生物化学性。他承认

神经蛋白可能不是宇宙中唯一能够承载精神生活的物质，正像除叶绿素之外，许多别的物质（或许在火星上）也可能催化碳水化合物的合成一样。但是他否认金属或硅作为替代物的潜在可能，在火星上也不例外。他问道：难道一台用旧啤酒罐制成的计算机有可能产生**理解**吗？对于这个夸张的问题只能响亮地回答："否！"简言之，塞尔认为，用来制造（当今的）计算机的无机物质，根本不可能承载心理功能，这在直觉上是显而易见的。

为了评价塞尔对计算心理学所作的两点尖锐的批评，我们先来看看他的这一观点：意向性必须以生物特性为基础。我们不妨称之为肯定性论断。与此成对照的是他的否定性论断：纯形式主义理论不能解释心理特性。然而，这里所作的假定超出了应有的程度，因为它的解释力是虚妄的。塞尔提出的生物学类比是一种误导，同时，他所求助于的直觉也是靠不住的。

塞尔告诉我们，大脑产生意向性可与光合作用相比较，情况真是这样吗？我们可给光合作用的**生成物**下定义，把碳水化合物总类中的各种糖类和淀粉清楚地区别开来，并说明它们与其他生物化学产物如蛋白质有何不同。此外，我们不仅**知道是**叶绿素支持着光合作用，同时也**理解**它是**怎样**完成这个过程的（以及**为什么**其他各种化学物质不能完成）。我们知道它只是一种催化剂，而不是原料；我们可以详细说明，它的催化作用在什么时刻，经由什么亚原子过程来完成。至于大脑和理解力，情况就全然不同了。

我们关于意向性是什么的理论（不管它是怎样产生的）是无法与我们关于碳水化合物的知识相比的，仅就意向性是什

么而言,在哲学上也还是有争议的。我们甚至不能完全有把握地说,看到它时就能将它识别出来。一个得到普遍认同的看法是,命题态度是意向性的,感觉和知觉则不是;但是,关于情感的意向性,尚缺乏清晰的一致意见。

为了表征意向性,并区分它的亚种作为不同的意向状态(信念、欲望、希望、意向及类似的东西),已经做过种种尝试。从塞尔早期关于言语行为的著作(Searle 1969),到近期对于意向性的总的论述(Searle 1983),他本人对此作出过许多重要贡献。(由布伦塔诺在19世纪采用,同时也被塞尔采用的)一个通用判据是**心理学**判据。用布伦塔诺的话来说,意向状态使心灵指向一个对象;用塞尔的话来说,它们具有内在的表述能力,或是"所指内容";无论何种情况,它们都把心灵同这个世界、乃至种种可能的世界联系起来。但是有一些作者是用**逻辑**术语来定义意向性的(Chisholm 1967)。至于逻辑定义和心理学定义是否精确地在范围上相对应,也不甚清楚(Boden 1970)。总之,没有一个意向性理论像关于碳水化合物的化学理论那样,是在没有问题的情况下被接受的。

至于大脑对意向性的生物化学"合成",就显得更加神秘了。我们有很充分的理由相信是这样:神经蛋白承载着意向性,但是作为神经蛋白,它是**怎样**具备这种能力的,我们却一无所知。

就我们对这些事情的理解而言,我们关注的是在神经元和突触中实现的某些**信息功能**(如信息传递、助长、抑制)的神经化学基础。例如,细胞壁上的钠泵怎样使一个动作势能沿着轴突传播开来;电化学变化怎样使神经元进入不应期,或从中恢复;或者神经元阈值怎样能被神经递质如乙酰胆碱改变。

例如，对于一个视觉细胞来说，一个至关重要的心理学问题可能是：**它是否能够具备检测强度梯度的功能**。如果神经生理学家能够告诉我们，哪些分子能完成这一任务，情况就会好得多。但是从心理学的观点来看，重要的并不是生物化学性本身，而是建立在它之上的担负信息的功能。〔塞尔显然同意这种观点，因为他说："对大脑中意向状态的实现方式类型所作的描述，可以在比有关神经元的特定生物化学性高得多的功能层次上进行。"(Searle 1983:272)〕

正如"计算机视觉"研究所表明的，金属和硅毫无疑问能够承载视觉中含有的二维至三维映射所要求的某些功能。此外，它们还能实现用作识别强度梯度的特定数学功能(即做高斯差分计算的"狗侦探")，这种功能似乎存在于许多生物视觉系统中。应该承认，可能金属和硅不能承载包含在正常视觉中或一般理解过程中的全部功能。也许只有神经蛋白才能做到这一点，所以只有作为"地球生物"一员的人，才能够享有意向性了。但是目前，并没有特别的理由支持这种想法。在这种背景下，最重要的是，即使将来我们有理由这样想，那也必须是建立在经验发现的基础上，直觉是无济于事的。

如果有人问，哪些心-物依存关系从直觉上看是合理的，答案必然是：**一个也没有**。没有一个对(与动作电位相对立的)意向性苦苦思索的人说过："钠——当然是它！"钠泵"显而易见"不合理的程度并不比硅片少，电极性"显而易见"不相关的程度也不比旧啤酒罐少，乙酰胆碱令人感到意外的程度几乎与啤酒一样。这三对中每对的前一部分在**科学上**都是令人信服的，但这种情况并没有使它们哪一个可从**直觉上**作出理解：我们最初都是感到惊异的。

我们的直觉也许会随着科学的进步而变化，可能有一天我们终将看到神经蛋白(也许还有硅)显然能实现心灵，正像我们现在明白一般生物化学物质(包括叶绿素)显然能产生出另一些这类物质一样，然而在尿素合成之前，即使对化学家来说，这种直觉也并非是显而易见的。无论怎样，眼下谈论意向性的物质基础时，我们的直觉是毫无用处的。塞尔的"肯定性"论断，即他假定的对意向性的一种替代解释，从最好处说，不过是一个承诺性的注释，从最坏处说，只是在兜售神秘主义而已。

塞尔的否定性论断，即形式计算理论不能对理解作出解释，迟迟未受到反驳。我这里的反驳包括两部分：第一，直接论及他的中文屋的例子；第二，涉及他的背景假定——计算机程序是纯句法的(这也是中文屋例子的基础)。

中文屋的例子，在认知科学圈内外都引起很多争论。有一些批评，塞尔自己在他的原文中已预见到，另一些则以同行评论的形式出现(同时有塞尔的答复)，更多的则发表在其后。我这里只着重谈两点：塞尔所谓的机器人应答，以及我所谓的英文应答。

机器人应答承认，塞尔例子中对中文仅有的理解是屋外中国人享有的理解，屋中的塞尔没有将中文字符同外部世界中的事件联系起来的能力，这表明他不理解中文。类似地，即使尚克的电传打字计算机能够正确"回答"我们关于餐馆的问题，这台不能辨认餐馆，不能向服务员交费，或是不能咀嚼食物的计算机，不会对餐馆作出任何理解。但是机器人又是另一回事，因为机器人不仅带有餐馆脚本，而且还配备有照相视觉程序，以及能走路和能拾起东西的肢体。如果这样一种机

器人的输入输出行为与人类行为完全等同，那么就可以证明它既理解餐馆，又理解人们与它交流用的自然语言——也可能是中文。

由于机器人应答承认了认知不仅是形式符号处理的问题，同时还需要增加一组与外部世界的因果关系，因此塞尔对该应答所作的第一个反应是声称已经取得了胜利，其次，他坚持认为给计算系统增加感知运动能力，并不是增加意向性或理解。

为了论证这一点，他想象出一个机器人，里面有一个小小的塞尔，可能是在头颅里，而不是配备着一个使它工作的计算机程序。机器人中的塞尔在一本(新)规则书的帮助下，把纸片重新排列，把“甲”和“乙”递进递出，正像在他之前的屋中塞尔所做的那样。但是现在进来的部分或全部中文字符不是中国人送入的，而是由机器人眼睛和耳朵里的照相机和听力装置通过因果过程激发产生的。递出的中文字符不是由中国人的手接收，而是由接在机器人肢体上的马达和杠杆接收，从而造成运动的结果。简言之，这个机器人显然不仅能用中文回答问题，还同样能够看东西、做事情：它能认出生的豆芽，并根据食谱要求，把豆芽扔进锅里，和我们所有其他人做的一样。

(上述计算机视觉方面的研究使我们看到，若要完成这个实际例子，中文词汇量还需大大扩充。AI 在语言加工方面的大量研究工作表明，在塞尔原来的“问题回答”例子中，为了表达规则，对于英文也存在着同样的要求。无论在哪种情况下，屋中的塞尔所需要的不是如此之多的中文，甚至也不是英文，而是编程语言。我们即将回到这一问题上。)

然而，机器人中的塞尔像他那位关在屋中的前任一样，对

更广泛的背景情况一无所知。他像原来一样对中文简直一窍不通，同时，对外部世界的掌握也不比前例中更多。对他来说，豆芽和锅既看不见，也抓不到；机器人中的塞尔所能看见和摸到的，除了规则书和涂划之外，就只有他自己的身体和机器人头颅的内壁了。塞尔说，因此我们不能认为机器人对这些实际事物有什么理解。其实，它根本看不到，也做不了任何事情：它只是"受电路和程序支配简单地来回运动而已"，其后由它里面的人例示说明，而这个人"没有任何相关类型的意向状态"。(Searle 1980:420)

塞尔在这里所作的论证不能作为对机器人应答的反驳而被接受，因为它在想象的例子与计算心理学的主张之间作了一个错误的类比。

根据塞尔的假定，机器人中的塞尔应实现人类大脑(根据计算理论)所实现的功能。虽然塞尔认为机器人中的塞尔应具有原模原样的意向性，就像他本人一样，但是其实大多数计算专家并没有把意向性赋予大脑(我们即将看到，那些把意向性赋予大脑的人，只是以非常有限的方式这样做的)。计算心理学并不相信大脑能**看见豆芽**或**理解英文**：像这样的意向状态是人的特性，而不是大脑的特性。虽然表述和心理过程(被计算专家和塞尔同样地)假定为在大脑中得到实现，但它们使之成为可能的感觉运动能力和命题态度是属于作为整体的人的。所以塞尔所说机器人颅内的系统是一个能理解英文的系统，并不是真的等同于计算专家们所说的有关大脑的情况。

的确，计算心理学家们所假设的、并由他们在心灵的计算机模型中实现的那些专门过程，是有点愚笨的，而且当人们朝更加基本的理论层次前进时，它们会显得越来越愚笨。作为

例子，我们来看看自然语言的语法分析理论。寻找一个限定词的语法分析过程，不是理解英文，为人称代词确定参照词的过程也同样不是理解英文，只有在大脑中完成这些解释过程的人，以及其他许多与这些过程发生联系的人们，才能够理解英文。理解英文的能力包含大量相互作用的信息加工过程，其中每一过程所完成的只是非常有限的功能，但是合在一起，它们就提供了以英文句子作为输入并以恰当的英文句子作为输出的能力。同样的看法也适用于视觉计算理论、问题求解或是学习的各个组成部分。正是因为心理学家希望**解释**人类语言、视觉、推理和学习，他们才设定了一些不存在这些能力的基本过程。

简言之，塞尔把机器人的假脑（也就是机器人中的塞尔）说成是理解英文的，这中间有一个范畴错误，相当于认为大脑是智能的担负者而不是智能的因果基础。

这里，有人也许会反对说，我是自相矛盾的：我主张不能把意向性赋予大脑，但是我的做法却隐含着这种倾向。因为我说过大脑实行"愚笨"的组合式过程，但是愚笨其实也是一种智能。愚笨就意味着有智能，但并非高智能（我们可以说一个人或一条鱼笨，但却不能说一块石头或一条河笨）。

我的辩护分作两步。第一，所有理论中最基本的理论层次是处在机器码的神经科学等价物中的，这是一个由进化"建造"的层次。某种感光细胞能用与"狗侦探"一样的动作对强度梯度作出响应，一个神经元可以抑制另一个神经元的激发，这些事实都可以用大脑的生物化学性来解释。在讨论这些事实的时候，愚笨的概念完全是不恰当的，即使是以引喻的方式。然而，这些很基本的信息加工功能（狗侦探、突触抑制）却

能正确地说成“非常、非常、非常……笨”。当然,这意味着,意向的语言,即使是属于十分牵强和不足以称道的类型,毕竟可以应用于大脑加工——由此引出我的第二点辩护。我不是说不能把意向性赋予大脑,而是说原来意义上的意向性不行。我也不是说大脑根本不能以任何有限的方式理解任何事物,而是说它们不能理解(例如)英文。几段之前,我甚至提到过,有些计算专家确实把某种程度的意向性赋予大脑(或赋予正在大脑中进行的计算过程)。让我们先来看一看英文应答及其与塞尔所作的“形式句法计算理论是纯句法的”这一背景假定的关系,这样,以上两点会更为清晰。

英文应答的关键问题是,例示计算机程序,无论是由人还是由人造的机器来完成,本身就包含着理解——至少是对规则书的理解。在塞尔原来的例子中,至关重要的一点就是屋中的塞尔能够理解书写规则所用的语言,即英文;同样,如果机器人中的塞尔不熟悉英文,机器人决不会把豆芽投进锅里。此外,如上所述,英文词汇量(对机器人中的塞尔来说,还有中文词汇量)必须经过大规模的修订,才能使这个例子生效。

我们可以把一种未知语言(无论是中文,还是B类线型文字)仅仅作为审美对象,或一组以系统方式联系起来的形式。逻辑学家或纯数学家能够仅仅从人工语言在心灵中的结构特性出发设计和研究人工语言〔虽然D·R·霍施塔特(Hofstadter 1979)的准算术pq系统的例子表明,形式演算在心理学上令人信服的、富有预见性的解释,可能会自发出现〕。然而,人们通常又用一种全然不同的方式对自己的母语符号作出响应。的确,要把熟悉的词的意义“括起来”(忽略掉),是很难做

到的。计算心理学家所持的观点,即自然语言可以用过程术语来表征,在这里是至关重要的,因为词、分句和句子可以看作微型程序。人们所理解的自然语言中的符号,引起了各种类型的心理活动。学习一种语言,就是建立起相关的因果联系,这不仅是词与现实世界("猫"与席子上的那个东西)之间的联系,而且是词与解释这些词时所具有的许多非内省过程之间的联系。

而且,我们毋需由(塞尔所作的)假设得知屋中的塞尔理解英文,因为他在屋中的行为清楚地表明他理解英文,或者更确切地说,他的行为表明他理解英文中一个**非常有限的子集**。

屋中的塞尔可能患有严重的记忆缺失症,致使他英文词汇中的百分之九十九被遗忘,但是这不会带来什么影响。他掌握了他所需的英文,这正是解释(原文如此)规则书所必须的,即详细说明怎样接收、选择、比较和送出不同模式的那部分。与塞尔不同的是,屋中的塞尔不需要像"催化"、"啤酒罐"、"叶绿素"和"餐馆"这样的词。但是他可能需要"找出"、"比较"、"二"、"三角形"和"窗口"这样的词(尽管他对这些词的理解可能比塞尔的完整理解差得多)。他必须理解条件句,因为规则也许这样规定:如果看到"甲",就应当送出"乙"。同样,他必须理解用某种方式表示的否定、时间顺序和概括(特别是他打算学会把工作干得更快时)。如果他使用的规则中包括某些分析中文句子的规则,那么他也需要语法范畴的词汇。(他不需要分析英文句子的明确规则,如 AI 程序中用于语言加工的分析过程那样,因为他已经理解英文。)

简单说,屋中的塞尔需要理解的仅仅是塞尔英文的子集,该子集等价于计算机所理解的编程语言,这台计算机会在窗

口生成同样的"问题回答"输入输出行为。类似地，机器人中的塞尔必须能够理解的无论何种英文子集，都等价于一台完全计算机化的、有视觉运动的机器人所理解的编程语言。

上述两点，可以说是把有争议的假定当成了论据。的确，像这样说计算机所理解的编程语言，看来是自相矛盾的。因为塞尔的基本前提(塞尔假定，这一前提是所有辩论参加者都接受的)是，计算机程序本质上是纯形式的：由它规定的计算是纯句法的，没有可供理解的内在意义或语义内容。

如果我们承认这一前提，上述的英文应答就可以立即勾消，因为这是在无对应比较可言的情况下强作比较。但是如果我们不承认该前提，如果——对不起塞尔〔还有其他人(Fodor 1980; Stich 1983)〕——计算机程序不仅与句法有关，那么英文应答毕竟可能是很有意义的。现在我们必须回过来论及这个基本问题。

当然，为了某种目的，人们可以把计算机程序看作一种未经解释的逻辑演算。例如，人们也许能够用纯形式的方法证明：一个形式得当的特殊公式，可以从程序的数据结构和推理规则中推导出来。而且确实有一个所谓的解释者程序，能够接受输入表结构"〔父亲(MAGGIE)〕"，并回答"(LEONARD)"[①]，它这样做时依据的只是形式判据，而无法把这些模式解释为可能是代表真实的人。同样，正如塞尔所指出的，配有餐馆脚本的程序，并不因此而具备有关餐馆的知识。形式体系与某一领域之间存在映射，其本身并不能为形式体系的处理者提供关于这一领域的任何理解。

---

① MAGGIE 和 LEONARD 分别是儿子和父亲的名字。——译者

但是决不要忘记,计算机程序是**为计算机设置的程序**,当程序在适合的硬件上运行时,机器总是要**做**某些事情的(所以计算机科学中使用“指令”和“执行”这两个词)。在机器码的层次上,程序对计算机的作用是直接的,因为机器的设计使得一个给定的指令只产生唯一的操作(高级语言的指令在执行前必须转换成机器码指令)。这样,程序指令就不仅仅是形式模式——甚至也不是说明性陈述(尽管为了某些目的,可以把它看作这些说法中的一种)。一旦给定适合的硬件背景,它就是一个能使当前步骤得以执行的过程说明。

对此,人们可能会说,编程语言是一个媒介,它不仅用来表达一些**表述**(一些可以写在纸上或装入计算机的结构,其中有些结构可能与人们感兴趣的事物保持同构),而且也用于使特定的机器产生**表述性活动**。

人们甚至会说,一个表述**是**一种活动,而不是一个结构。许多哲学家和心理学家都认为心理表述是内在能动的。最近有些人发表了这种见解,霍施塔特(Hofstadter 1985:648)便是其中之一,他特别批评了纽厄尔把**符号**作为可处理的形式标记的观点。按照他的说法,“大脑本身并不‘处理符号’,大脑是一个媒介,符号在它里面浮动着,并在它里面相互触发。”与“形式主义”心理学理论相比,霍施塔特对“联结论”心理学理论的热情更高。联结论方法含有一些很容易使人联想起大脑的并行处理系统,并且很适合于对大脑的表象、符号或概念建立**动态**模型。但是,不仅是联结论者能把概念看作内在能动的,也不仅是**大脑**表象可以用这种方式来看待,事实上,该主张已推广到传统计算机程序方面,特别是冯·诺伊曼机的设计上。计算机科学家 B·C·史密斯(Smith 1982)论证说,编程形

式的表述也是固有能动的，并且有关编程语言语义学的适当理论是承认这一事实的。

史密斯指出，当前计算机科学家们对这些问题有一种极为不适当的理解。他提醒我们，正如上面所说，关于**意向性**是什么，无论在计算机科学圈内还是圈外，都没有达成一致的看法，同时，关于**表述**也存在着深层的模糊性。即使从更为技术性的层面来讲，根据**计算**和**形式符号处理**，模糊性同样是无法避免的。因为计算机科学家对这些现象究竟是什么的理解，在很大程度上仍然是依靠直觉。史密斯关于编程语言的讨论，明确指出了计算机科学内部的某些根本性混乱。其中尤为重要的是，他认为，计算机科学家们通常在程序的控制功能与它作为形式句法系统的性质之间，作了过分的理论分离。

史密斯批评的这种理论割裂，在广泛使用的"双重演算"编程方法中，表现得十分明显。双重演算方法在程序运行时的说明性（或指称性）表述结构与对它作出解释的过程语言之间，设立了截然的理论区分。的确，知识表述和解释程序有时候是用两种十分不同的形式体系来书写的（如分别用谓词演算和 LISP 语言）。然而，它们也常常用同一形式体系来表达。例如，LISP（即表处理语言的首字母缩写）就允许用形式上类似的方式表达事实和过程，PROLOG（即逻辑编程的缩写）也是如此。在这些情况下，双重演算方法规定，所涉及的（单一）编程语言，可以在理论上用两种十分不同的方式来说明。

为了说明当前争论的这种区分，假定我们希望得到一个家庭关系的表述，该表述可以用来为这类问题提供答案。我们可能决定运用表结构来表述像 Leonard 是 Maggie 父亲这样的事实。或者，我们可能更喜欢基于框架的表述，在这种表述

中,可以同时用“LEONARD”和“MAGGIE”来填充父亲框架中相关的姓名槽。此外,我们也可以选择一个谓词演算公式,认为存在两个人(即 Leonard 和 Maggie),并且 Leonard 是 Maggie 的父亲。最后,我们还可以使用英文句子“Leonard is the father of Maggie (Leonard 是 Maggie 的父亲)”。

这四种表述,每一种都可以写在或画在纸上(就像屋中的塞尔所用规则书中的规则那样),以便我们学会怎样处理相关的标记法之后,就能作出解释。另外,它们也可以在计算机的数据库中得到实现。但是要使它们能为计算机所用,必须有一个解释程序,例如,当我们“问”该程序“谁是 Maggie 的父亲”时,它就能够找到条目“LEONARD”。有点头脑的人,都不会在计算机还尚未具备**表处理**能力的情况下,就让计算机实现表结构;也不会在没有**填充槽**机制的情况下,将框架给计算机,在没有**推理规则**的情况下,向计算机提供逻辑公式,或者在没有**语法分析过程**的情况下,给计算机英文句子。(同样,人们知道塞尔不懂葡萄牙文,就不会给屋中的塞尔一本葡萄牙文的规则书,除非他们准备先教他学习这种语言。)

史密斯并不否认,在表达式的**指称输入**(广义地说,就是所能映射到它上面的实际或可能世界)与它的**过程结果**(广义地说,就是它所做或使之发生的事情)之间有着重要的差别。表达式“〔父亲(MAGGIE)〕”是与两个实际人物之间特定的亲代关系同构的(所以才可能被我们映射到这种关系上去),这个事实是一回事。表达式“〔父亲(MAGGIE)〕”可以使某台计算机确定“LEONARD”的位置,这个事实完全是另一回事。假如情况不是这样,就不会形成双重演算方法。但是史密斯认为,与其固守双重演算方法,还不如采用一个“统一的”编程语

言理论，通过设计，使它既包含指称方面，也包含过程方面，这样就会更加简练和减少混乱。

他指出，被双重演算分开的每一边的许多基本术语，既有重要差异，也有深层的理论共性。以**变量**观念为例，在对它的理解上，逻辑学家和计算机科学家有某种相似之处，他们都承认，一个变量可以在不同的时刻被赋予不同的**数值**。既然是这样，要就"变量是什么"建立两个不同的理论，就是多余的。然而，在某种程度上，逻辑学家和计算机科学家又把这一术语理解为不同的事物：例如，在 LISP 编程语言中，一个变量的值就是另一个 LISP 表达式，而在逻辑中一个变量的值，通常是指相对于形式体系本身的某个外部对象。这些差异应予以澄清——不只是为了避免一个系统试图**用**变量**对**变量作推理时出现的混淆。简言之，我们需要单一的"变量"定义，既容许它（在逻辑中）作说明使用，也容许它（在编程中）作过程使用。史密斯证明了类似的看法也适用于其他基本计算术语，然后，他概括出 LISP 语言语义的单一解释，并且描述了一种新的演算（MANTIQ），是以心灵中的统一方法设计出来的。

正如用变量对变量作推理的例子所表明的，一个统一的计算理论，可以阐释**反映型**知识是如何可能的。因为一旦得出这样的理论，一个系统对数据和过程（包括系统本身的内部过程）的表述基本上就是可比较的。这一理论上的优点在心理学上具有重大意义（也是史密斯进行研究的主要动力）。

然而就我们当前的目的而言，关键的一点是，**程序**和**计算**的基本理论应当确认计算机程序的基本功能是要使得一些事情发生。一方面可以认为符号逻辑只不过是在摆弄一些未经解释的形式演算（诸如谓词演算），另一方面也可以把计算逻

辑看作是在用数学形式规定的“机器”(诸如图灵机)中进行与时间无关的抽象关系的研究,但是这些方式之中没有一种能够用来对计算机科学作出恰当的描述。

根据史密斯的看法,那种常见的把计算机程序表征为完全句法的而非语义的作法,是错误的。任何计算机程序固有的过程结果,都给了程序一个语义的立足点,这里所说的语义不是指称性的,而是因果性的。这是与屋中的塞尔理解英文的类比,而不是与他理解中文的类比。

同样的看法也见于 A·斯洛曼(Sloman 1986a;1986b)的论述,他认为,程序指令和计算机符号必须在一定意义上看作是具有某些语义的,不管这语义受到何种限制。在因果性的语义中,一个符号的意义(无论简单还是复杂)是要通过参照它与其他现象的因果联系来寻找的。核心问题是“什么是该符号被建立和(或)激活的原因?”和“它的结果是什么?”答案有时提及观察者看得到的外部对象和事件,有时则不。

如果这个系统是人、动物或机器人,它可能具有能使它参照餐馆和豆芽的因果能力(这里,因参照外部对象包括不能观察的外部对象而引起的哲学上的复杂性可以忽略不计,然而斯洛曼对此所作的讨论是有益的)。但是,无论所说的信息加工系统是什么样的,这些回答有时会描述纯粹**内部**的计算过程——从而使其他符号得以建立,使其他指令得以激活。这些例子有用英文词汇揭示的屋中塞尔的内心解释过程(或许可以与为自发的自然语言加工而定义的语法分析和语义过程相比较),还有尚克文本分析程序中的计算过程。尽管这种程序不能用“餐馆”符号表示**餐馆**的意义(因为它同餐馆、食物之类的东西没有因果联系),但是它的内部符号和过程确实实现

了对另一些特定事物的某种最低限度的理解，例如对两个形式结构进行比较的意义。

人们可能感到，这种例子中包含的“理解”是**如此**之少，以致这个词根本不应该使用。这样也罢。斯洛曼说得很清楚，重要的问题不是“**一台机器何时理解了某件事情**?”(这个问题暗示存在着某个明确的断点，理解在那里终止了，这是一种误导)，而是“**为了能够作出理解，一台机器(无论是不是生物的)必须能够做到哪些事情**?”这个问题不仅关系到计算心理学是否**可能**，而且关系到它的**内涵**。

总的说，本文的论述证明了，塞尔对计算心理学的非难缺乏充分的根据。把屋中的塞尔看作计算机程序的例示，并不等于说他一点理解力也没有。既然形式主义计算心理学的理论应当比作计算机程序而不是比作形式逻辑，那么在原理上，计算心理学就并非不能解释意义是怎样附着于心理过程的。

参考书目

Boden, M. A (1970). 'Intentionality and Physical Systems.' *Philosophy of Science* 37: 200–14.

Chisholm, R. M. (1967). 'Intentionality.' In P. Edwards (ed.), *The Encyclopedia of Philosophy*. Vol. IV, pp. 201–4. New York: Macmillan.

Dennett, D. C. (1971). 'Intentional Systems.' *J. Philosophy* 68: 87–106. Repr. in D. C. Dennett, *Brainstorms: Philosophical Essays on Mind and Psychology*, pp. 3–22. Cambridge, Mass.: MIT Press, 1978.

Fodor, J. A. (1980). 'Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology.' *Behavioral and Brain Sciences* 3: 63–110. Repr. in J. A. Fodor, *Representations: Philosophical Essays on the Foundations of Cognitive Science*, pp. 225–56. Brighton: Harvester Press, 1981.

Hofstadter, D. R. (1979). *Godel, Escher, Bach: An Eternal Golden Braid*. New York: Basic Books.

—— (1985). 'Waking Up from the Boolean Dream; Or, Subcognition as Computation.' In D. R. Hofstadter, *Metamagical Themas: Questing for the Essence of Mind and Pattern*, pp. 631–65. New York: Viking.

Newell, A. (1980). 'Physical Symbol Systems.' *Cognitive Science* 4: 135–83.

Schank, R. C., and Abelson, R. P. (1977). *Scripts, Plans, Goals, and Understanding*. Hillsdale, NJ: Erlbaum.

Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

—— (1980). 'Minds, Brains, and Programs.' *Behavioral and Brain Sciences* 3: 417–24.

—— (1983). *Intentionality: An Essay in the Philosophy of Mind*. Cambridge: Cambridge University Press.

Sloman, A. (1986*a*). 'Reference Without Causal Links.' In B. du Boulay and L. J. Steels (eds.), *Seventh European Conference on Artificial Intelligence*, pp. 369–81. Amsterdam: North-Holland.

—— (1986*b*). 'What Sorts of Machines Can Understand the Symbols They Use?' *Proc. Aristotelian Soc.* Supp. 60: 61–80.

Smith, B. C. (1982). *Reflection and Semantics in a Procedural Language*. Cambridge, Mass.: MIT Ph.D. dissertation and Technical Report LCS/TR-272.

Stich, S. C. (1983). *From Folk Psychology to Cognitive Science: The Case Against Belief*. Cambridge, Mass.: MIT Press/Bradford Books.



作为经验探索的计算机科学：符号和搜索

A·纽厄尔和 H·A·西蒙*

计算机科学是对围绕计算机的现象所作的研究。这一领域的创立者们深知这一点，所以他们自称为计算机器协会。机器——不仅是指硬件，而且是指带程序的活生生的机器——就是我们所研究的有机体。

这是第十次图灵讲座。在我们之前，已经有九位人物在这一讲座上就计算机科学提出了九种不同的观点。因为我们的有机体，即机器，可以在多种层次上并从多种角度进行研究。我们今天在这里提出另一种观点，深感荣幸，这一观点已经渗透在我们因之而受表彰的科学研究中。我们想把计算机科学称为经验探索。

我们的观点只是诸多观点中的一种，以前的讲座可以说明这一点。然而，即使将这些讲座汇总起来，也并不能涵盖这门科学的整个范围。这门科学的许多基本方面尚未在这十次荣誉讲座中提及。当罗盘转动了一圈，计算机科学的每个方面都已讨论过的时候，如果这一天到来——当然这不会太快，那么新一轮循环又会开始。讲演者必须像兔子那样，年复一年地追逐那些一小步一小步增长的积累——在持续不断

的比赛中乌龟所取得的科学和技术的进步。每一年都有新的差距出现，都在召唤新的冲刺，因为在科学中是没有终点的。

计算机科学是一门经验学科。我们也可以把它叫作实验科学，但是像天文学、经济学和地质学一样，它的某些独特的观察方式和实验方式已跳出实验方法狭隘的旧框框。然而，它们是实验。每制造一台新的机器都是一次实验。事实上，建造机器是对自然界的提问，同时我们也通过对运作中的机器进行观察和用一切有效的分析和测量手段对它进行分析，来听取答案。每编制一个新程序都是一次实验，它是向自然界提出一个问题，而它的行为为获得答案提供了线索。无论机器还是程序都不是黑盒子，而是已经设计出来的人工制品，既有硬件也有软件，我们可以打开它们，朝里看一看。我们可以把它们的结构同它们的行为联系起来，并从单个的实验中得到许多有用的东西。我们不必制造 100 个比如说定理证明程序的复制品，用统计学方式证明它尚未克服因按期望的方式进行搜索而出现的组合激增。借助于一些运行过程来查看程序，揭示出它的缺陷，我们就可进行下一个尝试。

我们制造计算机和程序的原因是多种多样的。我们把它们制造出来以服务于社会，并作为完成社会经济任务的工具。

---

* A·纽厄尔和 H·A·西蒙，“作为经验探索的计算机科学：符号和搜索”，第十次图灵讲座，首次发表于《计算机协会通讯》19(1976.3)。版权属 1976 计算机协会公司。经允许重印。

艾伦·纽厄尔(A. Newell)，卡内基-梅隆大学，心理学和计算机科学教授。

赫伯特·西蒙(H. A. Simon)，卡内基-梅隆大学，心理学和计算机科学教授。

但是,作为从事基础研究的科学家,我们制造机器和编制程序是一种发现新现象和分析我们已知现象的方法。公众常常会对这一点有误解,认为建造计算机和程序只是因为它们能产生出经济用途(或是朝着这种用途发展系列中的中间环节)。我们应当看到,围绕计算机的现象有其深层意义,同时又尚未分明,需要大量的实验来揭示其本质。我们应当看到,像在任何一门科学中那样,从这样一些实验和理解中积累起来的收获会在长久的新技术获取中收到成效,而正是这些技术,将创造出有助于社会实现其目标的工具。

然而,我们这里的目标并不是寻求从外部世界作出理解。我们是要考察这门科学的一个方面——通过经验探索而形成的新的基本理解。对此,最好是通过实例来说明。如果在这一场合我们从自己的研究领域中选取一些例子,我们会得到原谅的。正如我们将会看到的那样,这些例子涉及整个的人工智能发展,特别是它早期的发展。它们所依据的远远超出我们个人的贡献。即使在我们直接作出贡献的那些地方,也是同其他人合作的结果。在我们的合作者中,特别要提到的是克利夫·萧,整个 50 年代后期那段令人兴奋的时期里,我们与他组成了一个三人小组。此外,在卡内基-梅隆大学,我们还与许许多多同事和学生一起工作过。

由于时间的限制,只能举出两个例子。第一个是符号系统观念的形成。第二个是启发式搜索观念的形成。对于理解信息是如何加工的,以及智能是如何获得的,这两个概念有其深刻的意义。然而它们尚未达到对整个人工智能范围作出全面说明的程度,尽管在我们看来,它们在展示该部分计算机科学的基础知识性质方面是有用的。

1. 符号和物理符号系统

已有的基础方面的计算机科学知识的成果之一，就是在较为基本的层次上对符号是什么作出了解释。这种解释是关于性质的科学命题，是由经验得出的，它的形成是长期的、渐进的。

符号是智能行动的根基，这无疑是人工智能最重要的论题。因此，它对于整个计算机科学来说，也是最重要的问题。因为所有信息都是为一些目的服务而由计算机加工的，而我们衡量一个系统的智能水平，是看它在面临任务环境所设置的种种变动、困难和复杂性时，达到规定目的的能力。当所完成的任务范围有限时，计算机科学在实现智能过程中的这一总的投入并不引人注目，因为这时可以准确地预见这一环境中的全部变动。当我们将计算机扩展到更为综合、复杂和知识密集型的任务中去时，亦即我们试图让它们作我们的代理者，能够独自处理自然界中的全部偶发事件时，它就变得较为醒目了。

我们对于系统具有智能行动的必备条件的理解，是慢慢清楚起来的。它是复合型的，因为任何单个的基本事物都不能说明智能的全部表现。正如不存在能通过自己的特殊性质表示生命实质的“生命原理”一样，也不存在任何“智能原理”。然而，没有简单明了的解决办法，并不意味着智能在构造上没有任何必备条件。一个这样的必备条件就是存储和处理符号的能力。为了提出这一科学问题，我们或可对 W·麦卡洛克

(McCulloch 1961)著名文章的标题作出释义:符号是什么,使得智能可以使用它?智能是什么,使得它可以使用符号?

定性结构定律

所有科学都刻划出它们所研究的系统的基本性质的特征。这些特征在性质上表现出稳定的定性特点,因为它们设定了说明方式,在这个说明方式中可以形成更为详细的知识。它们的实质常常可以通过非常简明、非常一般的陈述来捕捉。如果不是出现了那些证明它们是最重大的成果的历史证据,人们也许会因为这些普遍定律的局限的特性,而断定它们对整个科学只作出了较小的贡献。

生物学中的细胞学说 生物学中的细胞学说是一个很好的定性结构定律的例子,它指出,一切活的有机体的基本组成单元是细胞。细胞以大量的不同形式出现,虽然它们都有一个核,周围是细胞质,而整个细胞由一层膜包围着。但是从历史上看,这种内部结构并不属于对细胞学说的说明,它是通过细致研究得出的后续特性。细胞学说可以由我们上面给出的陈述接近完整地表达,至于细胞的大小如何,则是一个模糊的观念。然而这一定律给生物学带来的影响是巨大的,在这一学说被逐渐接受之前,该领域丧失的活力是相当可观的。

地质学中的板块构造说 地质学为定性结构定律提供了一个令人感兴趣的例子。它之所以令人感兴趣,是因为它是近十年间才得到承认的,所以它变得举足轻重的情况,我们还记忆犹新。板块构造理论认为,地球表面是由一些大板块拼集而成的,这些板块共有数十块之多,它们(以地质学的速率)

运动着，相互背离、重叠，或是向下进入地心，到了地心，就失去它们的原貌。板块运动对于大陆和海洋的形状以及相对位置作出了解释，并解释了火山和地震活动的区域，以及深海海脊等等。增加一些速率和大小方面的细节，这一重要理论就得到详细说明了。当然，在成功地解释了许多细节之后，它才得到承认，这些细节全都一致符合（例如对西非与南美东北部之间植物群、动物群和地层一致性的解释）。板块理论表现出高度的定性特点。目前它得到了承认，整个地球似乎到处都在为它提供证据，因为我们是按照它去认识世界的。

细菌致病理论　自巴斯德阐明细菌致病理论以来，已过了一个多世纪，这个理论是在医学界引起一场革命的定性结构定律。这个理论指出，大多数疾病是由身体中微小的、活的单细胞有机体的存在和繁殖引起的，疾病的传染就是这些有机体从一个宿主传播到另一宿主。该理论的大部分精心研究在于确定出与特定疾病相联系的有机体，对它们加以描述，并追踪它们的生存过程。这个定律有许多例外情况——有许多疾病不是由细菌引起的，但这一事实并没有降低它的重要性。这个定律告诉我们要寻找特殊的致病原因；且并不认为我们总会找到这种原因。

原子论学说　原子论学说与上面刚提到的三个定性结构定律形成了有趣的对照。因为它出自道尔顿的研究工作和他对化学品以固定比例相结合的证明，所以这定律提供了定性结构的一个典型例子：元素是由细小的、均匀的颗粒组成的，一种元素的颗粒不同于另一种元素。但是由于原子的基本种类是如此简单，其变化范围是如此有限，所以很快形成了定量理论，它们吸收了原来定性假设中的所有一般性结构。从细胞、构造

板块和细菌可见，结构的种类是如此之多，这说明基本的定性原理具有特殊作用，它对整个理论的贡献是清晰可见的。

结论 定性结构定律在科学中随处可见。我们可以从中看到某些人类最伟大的科学发现。正如这些例子所表明的，它们往往建立了整个科学赖以运作的说明方式。

物理符号系统

让我们回到符号论题上，对**物理符号系统**作出定义。形容词"物理的"指明两个重要特征：

1. 这种系统显然是遵循物理学定律的，它们可由用工程化分量构成的工程化系统来实现；

2. 虽然我们使用术语"符号"，预先勾画的是我们意向式的解释，但是它并不局限于人类符号系统。

一个物理符号系统是由一组叫做符号的实体组成的，这些实体是一些物理模式，可以作为另一种叫做表达式（或符号结构）的实体的分量而存在。所以一个符号结构是由一些以某种物理方式相联系的符号实例（或标记）组成的（如一个标记紧接着另一个标记）。在任一瞬间，该系统都包含一个这些符号结构的集合体。除了这些结构而外，该系统还包含一个由按照一些表达式运作，以产生出另一些表达式的过程，如创造过程、修正过程、再生过程和破坏过程组成的集合体。物理符号系统是一架机器，它产生出一个随时间而演化发展的符号结构集合体。这种系统存在于一个对象世界之中，这些对象的数量比符号表达式本身更多。

对这些表达式、符号和对象的结构而言，有两个核心观

念:指称和解释。

指称　一个表达式指称一个对象是指,在已知该表达式的情况下,一个系统或是能够对该对象本身施加影响,或是能够以取决于该对象的方式规范其行为。

在每一情况下,其结果都是经由表达式而抵达对象,这正是指称的实质所在。

解释　系统能够解释一个表达式是指,该表达式指称一个过程,同时在已知该表达式的情况下,系统能够执行这一过程。

解释意指依赖行动的特定形式:已知一个表达式,系统就可以完成所指定的过程,也就是说,它能根据指称这些过程的表达式而再现和执行它所拥有的过程。

根据上述理解,一个具备指称和解释能力的系统必然也满足若干新增的必备条件,具有完备性和封闭性。由于篇幅的关系,我们只能简单谈这些。所有这些都是重要的,而且有着十分深远的影响。

1. 符号可以用来指称任何一种表达式。就是说,给出一个符号时,并没有事先规定它能指称什么表达式。这种任意性仅仅是对符号而言的,而符号标记和它们的相互关系则决定着一个复杂表达式所指称的是什么对象。

2. 存在着一些表达式是指称计算机所能完成的每一过程的。

3. 存在着一些过程是以任意方式建立任何表达式并修正任何表达式的。

4. 表达式具有稳定性,它们一旦建立,在明确地被修改或被取消之前,会一直存在下去。

5. 系统所能拥有的表达式个数基本上是无限制的。

我们刚刚定义的这种系统是计算机科学家们所熟悉的。从其属性来看,它与一切通用计算机极为相似。如果采用符号处理语言,如 LISP 语言,来定义一台机器,那么其亲缘关系就真的变得像同胞一样了。我们策划这种系统的意图并不是要提出某种新东西,恰恰相反,我们是要表明,对于符合这种特征的系统,我们现在已知什么,又作出了什么假设。

现在我们可以阐述一个一般性的科学假设了——符号系统的定性结构定律:

物理符号系统假设 对一般智能行动来说,物理符号系统具有必要的和充分的手段。

所谓"必要的"是指,任何表现出一般智能的系统都可以经分析证明是一个物理符号系统。所谓"充分的"是指,任何足够大的物理符号系统都可以通过进一步的组织而表现出一般智能。我们想用"一般智能行动"来表示与我们所看到的人类行动范围相同的智能:在任一真实情境中,对该系统目的来说是恰当的、并与环境要求相适应的行为,会在一定的速率和复杂性的限度之内发生。

物理符号系统假设显然是一个定性结构定律,它规定了系统的一般类别,在这些系统中我们会看到那些具有智能行动能力的系统。

这是一个经验假设。我们已定义了一个系统类别,还希望了解这个类别的系统是否说明了我们在现实世界中看到的一组现象。在我们周围的生物界中,主要是在人类行为中,智能行动随处可见。它是一种行为形式,我们可以根据其结果识别它是否是由人类完成的。该假设的确可能有误。智能行为并非这么容易产生,以至于任何系统不管其愿意不愿意都会表现

出这种行为。有些人经过分析,确定能以哲学或科学为根据得出这一假设为误的结论。而科学的态度是,只有拿出关于自然界的经验证据,我们才能攻击这一假设,或为它作出辩护。

现在我们需要回顾一下这一假设的形成过程,并看一看它的证据。

符号系统假设的形成

物理符号系统是通用机的一个例子。所以符号系统假设就意味着智能将由一台通用的计算机来实现。然而这一假设远远超出了通常在物理决定论的一般基础上得出的论点:任何可实现的计算,只要得到详细说明,就可以由通用机来实现。由于它明确地断言智能机器是一个符号系统,所以它以特定的构造方式断言了智能系统的性质。了解这一新增的特性是如何出现的,十分重要。

形式逻辑 该假设的起源要追溯到弗雷格、怀特海和罗素就形式化逻辑提出的方案:以逻辑方式获取基本的概念式数学观念,把证明和演绎观念置于可靠的根基上。这种努力在数理逻辑中达到了顶点,这就是我们所熟悉的命题逻辑,一阶和高阶逻辑。它形成了一种独特的观点,常常被称为"符号游戏"。逻辑,也包括所有数学,是根据特定的纯句法规则,用无意义的标记所做的一场游戏。所有的意义都被清除了。我们具有的是一个机械系统,尽管是非强制性的(我们现在也称之为非决定论的),而有关这一系统的种种事情都可加以证明。这样的成功首先是通过一步步远离所有看来与意义和人类的符号相关的东西而取得的。我们可以把这一阶段称为形

式符号处理阶段。

这种一般性态度充分反映在信息论的发展中。申农曾定义了一个仅仅对通信和选择有用的系统,而与意义毫无关系,这一点一再被指出。人们后悔以“信息论”这个一般性名字为这一领域命名,而打算把它修改为“选择信息的理论”,当然没有成功。

图灵机和数字计算机 早期数字计算机的发展和自动装置理论的发展可以放在一起讨论,它们是以图灵本人 30 年代的研究为起点的。在何者是根本的这一点上,它们的观点是一致的。我们采用图灵本人的模型,因为它充分表明了有关特点。

图灵机由两种存储装置组成:无限长纸带和有限状态控制装置。纸带上有数据,即人所共知的 0 和 1。该机器在纸带上有非常小的一组适当的操作——读、写和扫描操作。读的操作不是数据操作,而是为控制状态提供条件分支,而控制状态则表现为读头下数据的函数。正如我们都知道的那样,就计算机能做什么而言,上述模型包含了一切计算机的基本要素,虽然别的带有不同存储装置和运作装置的计算机也可能以不同的空间和时间条件完成同样的计算。特别要指出的是,图灵机模型内含两种观念:关于不能计算的东西的观念,以及关于通用机——即能做任何机器所能做到的任何事情的计算机——的观念。

30 年代,在现代计算机问世之前,我们已在两个方面对信息加工取得了深刻的认识,这的确是令人惊讶的。它对 A·图灵的天才创造具有启迪作用,并对当时的数理逻辑发展作出了贡献,计算机科学也无疑深深受惠于它。与图灵的著作

同时,还出现了逻辑学家E·波斯特和A·丘奇的(各自独立完成的)著作。他们从独立的逻辑系统观念(分别是波斯特生成和递归函数)开始,在不可判定性和通用性方面得出了类似的结果,这些结果不久即得到证明,原来这三个系统完全是等价的。的确,所有这些定义最一般的信息加工系统类别的尝试取得了一致的结论,这种情况具有某种说服力,使我们确信:我们已在这些模型中获得了信息加工的基本要素。

从表面上看,这些系统中没有一个把符号概念当作某种作指称用的东西。数据仅仅被看作一些0和1的数字串——对于计算向物理过程的还原来说,数据无作用这一点的确非常重要。有限状态控制系统总是被看作一个小控制器,为了在不破坏机器通用性的情况下弄清小到何种程度的状态系统是可用的,得做一些逻辑游戏。就我们所知,未曾有一种游戏为有限控制增添了新的动态方式的状态——把控制存储看作是持有大部分该系统的知识。在这一阶段上所完成的只是解释原理的一半——证明机器可以根据说明运转。因而,这是自动形式符号处理的阶段。

存储程序概念 在40年代中期,(在电子数字积分计算机之后)随着第二代电子计算机器的发展,出现了存储程序概念。将这一进展誉为在概念和实践两方面的里程碑是恰如其分的。现在程序可以是数据,也可以作为数据来运作。当然,图灵的模型中已隐含着这种能力:说明与数据同在一条纸带上。然而只有在机器获得足够的内存,使得在某个内部位置上找到实际程序在实践上是可行时,这一思想才得以实现。电子数字积分计算机毕竟只有20个字节。

存储程序概念实现了解释原理的另一半,这一部分指出

可以对系统拥有的数据作出解释。但是它仍未包含指称的观念——作为意义基础的物理关系的观念。

表处理 下一步——表处理,是在1956年完成的。现在数据结构内容成了我们物理符号系统意义上的符号,即被指称的模式意义上的符号,因而具有了所指对象。一些表拥有容许向另一些表做存取的地址——这就是表结构的观念。在表处理刚出现时,同行们一再向我们提出数据存在于何处的问题,就是说,最终是哪一个表拥有作为系统内容比特的集合体,这种情况向我们证明这是一个新观点。同行们惊异地发现,根本没有这样的比特,有的只是指称别的符号结构的符号而已。

在计算机科学的发展中,表处理同时表现为三件事情:

1. 它是机器中真正的动态存储结构的创建,而以前一直认为机器只有固定结构。它在替换和改变内容的操作之外,为我们的操作总体增添了建立和修正结构的操作。

2. 它及早地证明了这一基本抽象方式:计算机是由一组数据类型和一组对这些数据类型来说是恰当的操作组成的。这样,计算系统就能运用任何一种对应用来说是恰当的数据类型,而不受处于基础地位的机器影响。

3. 表处理产生出一个指称模型,而这样定义符号处理与我们今天把这一概念用于计算机科学时具有同样的意义。

正如经常出现的那样,当时的做法已预示出表处理的所有基本因素:地址显然被用于实现存取,磁鼓机则用于被连接的程序(所谓1加1编址),等等。但是,以抽象方式出现的表处理的概念形式开创了一个新天地,在这里,指称和动态符号结构定义出许多特征。将早期表处理系统嵌入语言(IPL、

LISP语言)的做法常常受到指责,说它是将表处理技术扩展到整个编程实践的障碍,然而它却是将抽象方式结合起来的工具。

LISP语言 还有一个步骤值得注意:麦卡锡在1959—1960年创立了LISP语言(McCarthy 1960)。它完成了对动作的抽象,是将表结构从它们在具体机器内的嵌入状态中取出,创建一个新的带有S表达式的形式系统。可以证明,它与其他通用的计算方案是等价的。

结论 指称符号和符号处理的概念直到50年代中期才出现,这并不是说较早的进展是非本质的,或不够重要。这一总的概念综合了可计算性、(通过多种技术的)物理可实现性、通用性、过程的符号表述(即可解释性),最后还有符号结构和指称。每一步进展都为这一整体提供了不可或缺的部分。

这个链条的第一个步骤是由图灵创立的,它是由理论方面的兴趣推动的,但是其他各步都深深地植根于经验。我们始终受到计算机本身进展的引导。存储程序原理出自电子数字积分计算机的经验。表处理方法出自构造智能程序的尝试,随机存取存储装置的出现使它受到启发,它为编址的指称符号提供了一个明显的物理实现方式。LISP语言则来自表处理方面不断发展的经验。

证　据

我们的假设是:物理符号系统具备智能行动的能力,同时一般智能行动也需要物理符号系统。现在我们来看看这一假设的证据。该假设是经验的概括,而不是定理。我们不知

道有什么方法可以在纯逻辑的基础上证明符号系统与智能之间的联系。既然缺乏这样的证明,就必须看一看事实了。然而我们的主要目的并不是审视这一证据的细节,而是运用我们掌握的例子来说明这一命题:计算机科学是一个经验探索的领域。因此,我们仅打算指明存在着何种证据,并指明检验过程的一般性质。

到50年代中期,物理符号系统观念已基本上具有目前的形式,我们还可以溯源至人工智能作为计算机科学直接分支的成长时代。从那时起,20年来的工作已展示出逐渐积累起来的经验证据,主要有两种类型。第一种着重说明物理符号系统对于产生智能的**充分性**,并试图构造和检验具有这种能力的特定系统。第二种证据着重说明,凡表现出智能的地方,具有物理符号系统的**必要性**。它以人为起点,因为这是我们最了解的智能系统,并试图弄清人的认知活动是否可以解释为物理符号系统的工作过程。还有另一些形式的证据,将在后面简单加以介绍,而上述两种证据特别重要,我们依次加以考察。前者一般被称为人工智能,后者一般被称为认知心理学研究。

构造智能系统　对细菌致病理论作初始检验的基本范式是:确定疾病,然后寻找细菌。人工智能研究深受类似范式的启发:确定需要智能的任务域,然后为数字计算机构造一个程序,使之能够处理这一任务域中的任务。一开始,我们看到的是一些简单的、精心构造的任务:难题和游戏、资源调度和分配的运筹学问题、简单的归纳任务。目前已构造出的这类程序,即使没有几百个,也有几十个之多,其中每一个都能在恰当的领域内完成某种程度的智能行动。

当然,智能不是一个全或无的事物,它在特定领域中不断地向更高级的性能发展,并向拓宽这些领域的范围发展。例如,早期的国际象棋程序只要能按规则走子,并表现出某种目的性,就被看作是成功的;稍后,它们达到了人类初学者的水平;在 10 到 15 年的时间里,它们开始对抗高水平的业余棋手。进步是缓慢的(同时在整个项目方面的投资也不多),但却从未间断,构造加检验的范式有规律地循环着——整个研究活动是在宏观层次上模仿许多 AI 程序的基本的生成加检验的循环过程。

我们看到,具备智能行动能力的领域在逐步拓宽。研究工作已从原来的任务扩展到建立系统:以各种方式处理和理解自然语言的系统、解释视觉场景的系统、用于手眼协调的系统、设计系统、编写计算机程序的系统、口语理解系统——这张单子如果不是无尽的,至少也是非常之长。即使存在着一些界限,在界限之外上述假设不再成立,这种界限也尚未分明。到目前为止,影响前进速度的主要是两个方面:已应用的科学资源的数量相当有限;对每一重大的新课题来说都不可避免地要为实际建立系统而付出努力。

当然,已着手研究的、适用于特定任务域的智能系统,远远不止上面罗列的这些例子。如果人们得知,完成这些形形色色的任务的 AI 程序,除了作为物理符号系统的实例而外,别无共同之处,可能会感到惊讶和沮丧。因而,对一般性所拥有的机制的搜索,以及对完成各种任务的程序中的共同成分的搜索,始终有其重要意义。这种搜索使该理论超越了原有的符号系统假设,以对人工智能中各种有效的特殊类型的符号系统的特征作出更加完美的说明。在本文第二节中,我们

将讨论一个处在后一说明层次上的假设的例子:启发式搜索假设。

对一般性的搜索衍生出一系列程序,这些程序是为了把一般问题求解机制从特殊任务域的必备条件中分离出来而设计的。一般问题求解程序(GPS)也许是第一个这样的程序,在它的后继者中则有像规划程序和协商程序这样的当代系统。对共同成分的搜索导致了表述目的和计划的一般化方案以及构造分辨网络的方法、控制树搜索的过程、模式匹配机制,以及语言分析系统的一般化方案。寻找一些方便的工具来表述时间和时态序列、运动、因果性,以及诸如此类的东西,目前正在进行这方面的实验。由这样的基本成分,以模件方式组装起一些大型智能系统,这种可能性变得越来越大。

我们回过来再看一看细菌理论的类比,可对目前的工作得出一些深入的看法。如果由细菌理论激起的第一次研究高潮主要在于找出伴随每一种疾病的细菌,那么接下来的工作就转为弄明白细菌是什么——在基本的定性规律的基础上建立一个新的结构层次。在人工智能中,最初旨在为范围广泛的、几乎是随机选出的各种不同类型的任务建立智能程序的活动高潮,正在让位于目标更集中的旨在理解这些系统的共同机制的研究。

建立人类符号行为模型 符号系统假设表明,人类之所以有符号行为,是因为人类具有物理符号系统的特征。因而,为带有符号系统的人类行为建立模型所作的努力,其结果就成了该假设证据的重要部分,同时,人工智能研究所作的工作也与通常称为信息加工心理学的研究紧密配合。

在过去的20年中,搜索根据符号系统对人类智能行为作

出解释的做法，已在很大程度上取得成功，达到信息加工理论成为认知心理学中当前的主导观点的地步。尤其是在问题求解、概念获取和长时记忆领域中，符号处理模型目前居于支配地位。

信息加工心理学研究包括两种主要的经验活动。第一种是对人类在完成需要智能的任务时的行为进行观察和实验。第二种与人工智能中的对应活动十分相似，即为了给观察到的人类行为建立模型而编制符号系统程序。心理学观察和实验导致对参试者正在使用的符号加工作出系统假设，这些假设就是那些参与构造程序的思想的重要源泉。因而，许多有关 GPS 基本机制的思想，就是对人类参试者在完成问题求解任务期间进行发声思考时所作报告进行仔细分析而得出的。

计算机科学的经验特性，无论在哪一方面，显然都未超过与心理学的这种亲缘关系。不仅在检验模拟模型是否确实可作为人类行为的解释方面，心理学实验是必不可少的，同时这些实验也是设计和构造物理符号系统的新思想的出处。

另外的证据 对我们现在已考察过的符号系统假设来说，其证据的主要方面来自反面：关于智能活动——无论由人还是由机器——究竟是怎样完成的，还没有可与之抗衡的专门假设。为建立这类假设所作的尝试大多发生在心理学领域之中。在心理学中，我们有一个从通常称为“行为主义”的观点到通常称为“格式塔理论”的观点的理论连续体系。这些观点中没有一个能成为符号系统假设的真正竞争者，理由有两点。第一，无论是行为主义还是格式塔理论都未曾证明，甚至未曾指出，如何证明它所假定的解释机制对于说明完成复杂任务的智能行为是充分的。第二，在对这两个理论的系统阐

述中,从未有过像人工程序那样的详细说明。事实上,这些可供选择的理论有些笼统,以至把信息加工解释赋予它们,从而使它们与符号系统假设相似,并不是极其困难的。

结　论

我们已经尝试以物理符号系统假设为例具体地说明计算机科学是一个科学事业,这里科学事业这一术语是按照它通常的意义使用的:它提出科学假设,然后寻求以经验探索来证实之。然而,我们选择这个特殊例子来说明我们的观点,还有第二个原因。物理符号系统假设本身就是我们前面称为"定性结构定律"的那种实际的科学假设。它代表计算机科学的一个重要发现,如果它由经验证据得到证明,实际情况正是如此,那么它将对这一领域产生出重要而持久的影响。

现在我们来看看第二个例子:搜索在智能中的作用。这个论题,以及我们将要检查的有关它的特殊假设,一般地在计算机科学中,特别地在人工智能中,都处于核心的地位。

2. 启发式搜索

知道物理符号系统为智能行动提供了基体,并没有告诉我们这些系统是如何做到这一点的。有关计算机科学中定性结构定律的第二个例子强调的正是后一问题,它指出符号系统是使用启发式搜索过程来解题的。与前面的情况一样,这一总结根据的是经验证据,而不是从其他前提中正式地推

导而出的。然而我们不久就会看到，它确实与符号系统假设有着某种逻辑联系，我们或许可以期望在未来某时将这种联系形式化。在此之前，我们的描述必然还是一种经验探索。我们将对已了解的有关启发式搜索的情况作一介绍，并回顾那些说明它如何能使行动成为智能行动的经验成果。我们就以说明这一定性结构定律——启发式搜索假设——作为开始。

启发式搜索假设 将问题的解表述为符号结构。物理符号系统在问题求解中以搜索方式行使它的智能，亦即生成符号结构，并逐步对其进行修正，直到产生出一种解的结构。

为了求解问题，物理符号系统必须使用启发式搜索，因为这种系统具有的加工资源是有限的——在数目有限的步骤中，同时在有限的时间区间内，它们只能执行数量有限的加工。当然，这不是一个十分强的限制，因为所有通用图灵机都会受到这种限制。然而，我们想说的是更强意义上的限制：我们意指的是**实际的**限制。我们可以构想这样的系统，它们不受实际的限制，如对一棵以指数方式展开的树的节点，在深度上以每单位进度速率相同的方式作并行搜索，但这是不可能的。这里我们关心的不是这种系统，而是这样一些系统，它们的计算资源相对于它们面临情境的复杂性来说，是稀缺的。这个限制对于真实任务语境中的、以计算机或人的形式出现的任何真实的符号系统都是存在的。对大多数目标来说，资源有限的事实使得我们将符号系统看成好像是一次完成一个加工的串行设备。如果它在任何一个短暂的时间区间内只能完成少量的加工，那么我们完全可以把它看作是以一次做一件事的方式在工作。这样，“有限资源符号系统”和“串行符号

系统"实际上就成为同义语。如果这个瞬间足够短，从瞬间到瞬间的稀缺资源分配问题，通常可被看作串行机的时间调度问题。

问 题 求 解

既然一般来说，解题的能力被看作系统具有智能的首要标志，那么人工智能的大部分历史自然也就与设法建立和理解问题求解系统有关。两千年来，问题求解一直是哲学家和心理学家讨论的对象，他们深奥的议论给人以神秘之感。如果你认为解题的符号系统没有什么疑问和神秘之处，那么你现在一定很年轻，因为你的观点是本世纪中叶才形成的。柏拉图(依他的说法还有苏格拉底)看到，即使理解问题是如何提出的也是很困难的，更何况理解它们如何可能被解决。请你回忆一下柏拉图在《美诺篇》中是如何提出这个难解之谜的：

> 美诺：苏格拉底，请问你是如何探讨你所不理解的东西的？你把什么作为探讨的主题？在你发现你想要的东西时，你如何知道这就是你以前所不知道的东西？

为了解决这个难题，柏拉图提出了他著名的回忆说：当你认为你是在发现或学习某个事物时，其实你不过是在回忆以前的生存中你已经知道的事物。如果你感到这种解释十分怪诞，那么今天可获得的、建立在我们对符号系统理解之上的解

释则简明得多了。它可近似地叙述如下：

说明一个问题就是指明：(1)对一类符号结构(该问题的解)的**检验**；以及(2)符号结构的**生成程序**(潜在的解)。解决问题就是运用(2)生成一个满足(1)的检验的结构。

如果我们知道我们想要做什么(即检验)，同时又不是即刻知道如何做到这一点(我们的生成程序没有即刻产生出一个满足这一检验的符号结构)，这时我们就得出一个问题。一个符号系统(有时)可以说明和解决问题，是因为它有生成和检验的能力。

如果问题求解要做的仅仅就是这些，为什么不立即简单地生成一个满足该检验的表达式呢？事实上，这正是我们梦寐以求的事情。“如果想要马就有马，那么乞丐都有马骑了。”当然，除非是在梦境中，这种事情不可能发生。一旦构造出某个事物，我们知道该如何检验它，这并不意味着我们知道怎样构造它，即并不意味着我们有任何一个能做到这一点的生成程序。

例如，大家都明白，“解决”赢棋问题意味着什么。表明获胜阵势的简单检验——将死对方的国王的检验，是存在的。在梦境中，人们只要生成一个达到将死对手所有对应策略的策略即可。可惜对于实际的符号系统(人或机器)来说，这样的生成程序是闻所未闻的。实际情况是，寻找下棋高招的方法是生成各种备选方案，通过采用近似的、常常是错误的标准，假定它们指出一条有可能通向获胜阵势的特定走棋路线，并不厌其烦地对这样的方案作出评价。移动生成程序是有的，但却没有获胜走步生成程序。

对一个问题来说，要有移动生成程序，必须先有一个问题

空间——一个符号结构空间，在这一空间中，问题的情境，包括初始情境和目标情境，都能够得到表述。移动生成程序是用来对问题空间的一个情境进行修正，使之进入另一情境的过程。物理符号系统的基本特征保证它们能够表述问题空间，并拥有移动生成程序。在任一具体情境下，这些系统怎样将问题空间和适合于这情境的移动生成程序结合起来，仍是一个在人工智能研究最前沿的问题。

因而，为符号系统提供问题和问题空间之后，符号系统所面临的任务就是使用它有限的加工资源去生成可能的解，一个接一个，直到找到一个解满足对所定义问题的检验。如果该系统对产生潜在解的顺序具有某种控制能力，那么就值得对这种生成顺序加以安排，从而使实际的解及早出现的可能性大大提高。如果一个符号系统在一定程度上成功地做到了这一点，它就有可能在这一程度上表现出智能。对一个加工资源有限的系统来说，智能就在于对下一步做什么作出明智的选择。

问题求解中的搜索

在人工智能研究第一个十年左右的时间里，问题求解研究与搜索过程研究几乎就是同义语。从我们对问题和问题求解特征的刻划来看，出现这种情况的原因是显而易见的。其实，也可以探问一下是否还有别的可能。但是在我们试图回答这一问题之前，必须对搜索过程在这十年活动期间自行显现的性质作进一步的深入了解。

从问题空间中抽取信息　我们来看一个符号结构集合，

它之中的某个小的子集合就是给定问题的解。我们进一步假定，这些解随机地分布在整个集合中。也就是说，不存在任何信息能使任一搜索生成程序比随机搜索干得更好。这样，在解题过程中，没有一个符号系统能比任何别的符号系统表现出更高(或更低)的智能，尽管某个符号系统的运气可能比另一个更好些。

因而，表现出智能的一个条件就是解的分布不是完全随机的，符号结构空间至少表现出一定程度的秩序和模式。第二个条件是，符号结构空间中的模式必须或多或少是可测出的。第三个条件是，潜在解的生成程序必须能够根据它测出的模式情况而区别其行为。问题空间中必然有信息存在，而符号系统必须能够抽取它和使用它。我们先来看一个非常简单的例子，在该例中智能很容易表现出来。

我们来看看这个求解简单代数方程的问题：

$$AX + B = CX + D$$

其检验是将解定义为任一形式如 $X = E$ 的表达式，使 $AE + B = CE + D$。当然，我们可以用无论什么过程作为生成程序，由该过程得出一些数字，然后将这些数字代入后一方程进行检验。我们不会把这种过程叫做智能生成程序。

换一种方式，我们可以采用一些利用这个事实的生成程序：在不改变原方程的解的情况下，可以对原方程进行修正——在两边加或减相等的量，或用同样的量乘或除两边。当然，我们还可以得到更多的指导生成程序的信息，其做法是将原表达式与解的形式加以比较，在方程中精确地作出使解保持不变的一些改变，与此同时使方程变成所希望的形式。这种生成程序会注意到原方程的右边有一个多余的 CX，将它

从两边减去，再将一些项合并。接下来它会注意到左边有一个多余的 B，于是减掉它。最后，它可以用除法将左边多余的系数(A－C)去掉。

这一过程表现出相当高的智能，这样，生成程序就通过它产生出相继的符号结构，每一结构都是通过对前一结构的修正获得的，而修正的目的是在保持解的其他条件不变的情况下，减少输入结构形式与检验表达形式之间的差异。

这个简单例子已经说明了不少以智能方式做问题求解的符号系统所采用的主要机制。首先，每个后继表达式都不是独立生成的，而是通过修正以前产生的表达式而产生的。第二，这些修正不是杂乱无章的，而是依赖于两种信息。它们依赖于在整个这类代数问题中保持不变的信息，而这些信息已被植入生成程序本身的结构中：对表达式的所有修正必须使方程的解保持不变。它们也依赖于每一步中的变化信息：测出当前表达式与期望表达式之间在形式上仍存在的差异。实际上，生成程序本身含有该解必须满足的一些检验，所以不满足这些检验的表达式是决不会生成的。对第一种信息的使用保证了在所有可能存在的表达式中，实际生成的只是一个很小的子集合，同时该子集合又不会遗漏解的表达式。使用第二种信息则是运用简单形式的手段目的分析方法得出搜索方向，通过一系列近似形式，达到所期望的解。

要问引导搜索的信息是从哪里来的，这没有什么神秘。我们不必追随柏拉图，把一个已经知道解的“以前的生存”赋予符号系统。一个适度复杂的生成程序检验系统不必求助于灵魂转世，就能达到目的。

搜索树 以简单的代数问题作为搜索的例子看起来也许

是不大常见的，甚至有点反常。它显然不是试错式搜索，因为虽然有一些试探，但却没有任何错误。我们更习惯于把问题求解搜索看作生成枝叉繁多的树，作为部分解，它们具有这种可能性：在得出一个解之前，会长出数以千计的、甚至数以百万计的分支来。这样，如果生成程序从它产生的每一个表达式中造出 B 个新的分支，那么这棵树就会长出 B^D 个分枝，D 是它的深度。对代数问题而言，这棵生长的树有其特殊性，那就是它的分支数 B 等于 1。

国际象棋程序一般会长成庞大的搜索树，在某些情况下，其分支数目多达百万以上。这个例子虽然可用来说明我们关于树搜索的观点，但是应该看到，国际象棋中的搜索目的不是生成拟议中的解，而是对解进行评价(作检验)。研究弈棋程序的路线之一，已集中在改进棋盘表述方式上，以及改进在其上走子的过程，以便提高搜索速度，使搜索更大的树成为可能。不难看出，该方向的基本原理为：动态搜索越深，对其结果的评价也就越准确。而另一方面，又有充分的经验证据表明：最高水平的人类棋手——大师们，极少开发出超过一百个分支的搜索树。这种经济性的取得，主要不是因为搜索的深度不及国际象棋程序，而是每一节点上的分支非常稀少，并且有选择性。在不致造成评价失真的情况下，要做到这一点，只能是将更多的选择性植入生成程序本身，这样，它就可以只选择生成那些极有可能提供对阵势来说有重要参考价值的信息的分支。

上述讨论造成了听上去有些自相矛盾的结论：搜索，亦即相继生成潜在解结构，是符号系统在问题求解中行使智能的基本方面，但是它所表现出的智能的多少，却不是由搜索的多

少来衡量的。一个问题成其为问题,并不是该问题的解需要进行大量的搜索,而是,如果不应用必要的智能水平的话,就需要进行大量的搜索。当致力于解决一个问题的符号系统对它要做的事情有充分的了解时,就会简捷地径直向它的目标前进;但是在进入未知领域时,它的知识一旦变得不够用了,在重新找到它的道路之前,会面临进行大量搜索的威胁。

在生成问题解答的每一个方案中,都存在着搜索树指数激增的可能性,这种情况警告我们,在补偿生成程序的无知和选择能力缺乏时不要依赖计算机的蛮力,哪怕是最大、最快的计算机。在某些人的胸中时而还会燃起这种希望:可以找到足够快的、并且充分巧妙地编程的计算机让它们通过蛮力搜索而高明地下棋。还未曾听说过在弈棋理论中能完全排除这种可能性。然而,关于规模相当大的树中的搜索处理的经验研究并未取得显著成果,这使我们看到,这一研究方向的前途远不如当初国际象棋首选为恰当的人工智能任务时显得那么光明。这一点应当看作是国际象棋程序研究方面取得的重要经验发现之一。

智能的形式 因而,智能的任务就是防止搜索中始终存在的由指数激增造成的威胁。这一点如何才能实现呢?第一种途径是将选择性植入生成程序,代数的例子已对此作出了说明,同时只为后继分析生成"似乎合理"走步的国际象棋程序也对此作出了说明,即只生成保证是解的、或是沿着通向解的路径的结构。这种做法的结果通常可以降低分支的速度,而不能完全杜绝分支。除了像代数例子那种例外地高结构化的情况而外,最终的指数激增是难以避免的,只不过是推迟它的到来而已。所以智能系统一般必须以别的使用信息引导搜

索的技术来增补它的解生成程序的选择性。

在处理各种各样任务环境中的树搜索方面已积累了20年的经验，这些经验已经产生出一小套一般性技术，它们成为当今每一位人工智能研究者的知识的组成部分。这些技术已在许多一般性著作中作过介绍，如尼尔森的著作（Nilsson 1971），所以我们可以在这里对它们作出十分简单的概括。

串行启发式搜索中的基本问题总是：下一步要做什么？在树搜索中该问题又分为两个方面：（1）下一步搜索从树的哪个节点开始，（2）从该节点起应取什么方向？有助于回答第一个问题的信息，可以解释为对不同节点与目标的相对距离的测定。佳者优先搜索要求从那个看来离目标最近的节点开始下一步搜索。获得有助于回答第二个问题的信息，即在什么方向上进行搜索，其方法正如代数例子所表明的那样，常常是测出当前节点结构与由对解的检验所描述的目标结构之间的特定差异，并选择那些对减少这些特殊类型差异有重要作用的行动。这就是称为手段目的分析的技术，它在一般问题求解程序的结构中起着主导作用。

经验研究作为AI研究中的一般思想源泉的重要性，可以通过回顾佳者优先搜索和手段目的分析这两个主导思想的历史，在大量问题求解程序中清楚地得到证明。佳者优先搜索的雏形在1955年的逻辑理论家程序中已经出现，虽然当时没有叫这个名字。体现了手段目的分析的一般问题求解程序，约在1957年出现，但是它是与改进的深度优先搜索相结合的，而不是与佳者优先搜索相结合。出于节约内存的目的，国际象棋程序一般是与深度优先搜索联姻的，大约在1958年之后，增补了强有力的α-β修剪过程。这些技术中的每一项都

数度东山再起，直到60年代中后期，一直很难看到根据这些概念对问题求解所作的一般性的、与任务分离的理论探讨。它们从数学理论中得到的形式依据，总计起来为数也不多：一些关于搜索还原的定理，这些定理可从使用 $\alpha-\beta$ 启发法中获得；两个关于最短路径搜索的定理（尼尔森曾提到过，Nilsson 1971）；以及最近才出现的一些带有概率论评价功能的佳者优先搜索定理。

“弱”方法与“强”方法　我们一直在讨论的这些技术都是用于控制指数扩张的，而不是阻止指数扩张。因此，把它们称做“弱”方法是恰如其分的——在问题空间实际含有的符号系统知识或结构数量不足以保证完全避免搜索时所使用的方法。对比一下高结构化情境与低结构化情境的不同是有益的，前者可以用公式表达为比如线性规划问题，后者则是像旅行售货员问题或时间调度问题那样的组合问题。（这里“低结构化”是指与问题空间结构相关的理论是不充分的或不存在的。）

在解决线性规划问题时，也许需要做数量可观的计算，但是这种搜索不出现分支，每一步都沿着通向解的路线行进。在解决组合问题或是证明定理时，树搜索就难以避免了，而成功则有赖于我们已描述过的那类启发式搜索方法。

AI问题求解研究的趋势并非全都依照我们以上概述的途径发展。对定理证明系统的研究就是一个观点有所不同的例子。在这里，从数学和逻辑引进的思想已对探索方向产生出强烈的影响。例如，在完备性得不到证明的情况下，就不能使用启发法（这有一点讽刺的意味，因为就我们所知，大多数有价值的数学系统是不可判定的）。由于对佳者优先搜索启

发法或对多种选择生成程序来说，完备性特性是很难得到证明的，这一必备条件就具有相当的约束力。当定理证明程序因它们的搜索树组合激增而一再显得无能为力的时候，人们开始把目光投向有选择性的启发法了，在很多情况下，这种方法原来不过是对一般问题求解程序中所使用的启发法的模拟。例如，支持集启发法就是一种适合于分解定理证明环境的反向工作形式。

经验总结 至此，我们已介绍了第二个定性结构定律的工作方式，这一定律表明物理符号系统是通过启发式搜索来解题的。除此之外，我们还考察了启发式搜索的某些附带特征，特别是它始终面临的来自搜索树指数激增的威胁，以及为了避开这种威胁而采用的某些手段。在有效的启发式搜索是怎样作为问题求解机制的问题上，看法产生了分歧，而看法的不同取决于涉及什么任务域，以及采纳什么样的恰当性判据。降低愿望水平可以使成功得到保证，反之，提高愿望水平则会招致失败。其证据也许可以大致归结如下：几乎没有什么程序是在“专家”的职业水平上解决问题的。最有名的例外情况要算萨谬尔的检查者程序和费根鲍姆及莱德伯格的 DENDRAL 系统了，但是我们也可以指出，有许多启发式搜索程序用于像时间调度和整体程序设计那样的运筹学问题领域。在许多领域中，程序的性能达到相当高的业余水平，如国际象棋、某些定理证明领域、类型众多的游戏和难题。具有复杂感知“前端”的程序，如视觉识别程序、口语理解程序、必须在实际时空中运动自如的机器人，距离人类水平还相当远。然而，在解决这些困难任务方面，已经取得了令人印象深刻的进步，并积累了大量的经验。

我们没有对那些已经出现的特殊性能模式作深入的理论解释。然而,我们可以在经验的基础上得出两点结论。第一,从我们已经掌握的人类专家在完成像国际象棋这样的任务时的表现来看,很有可能是,任何可与这种表现相匹敌的系统,都必须存取其存储器中存储的大量语义信息。第二,人类完成任务时在丰富的感知组成方面的优势,可部分地归因于人类眼和耳的那种生而有之的特殊目的的并行处理结构。

总之,性能质量必然依赖于问题的领域和处理它们时所用符号系统这两方面的特征。就我们感兴趣的大多数实际生活领域而言,至今尚未证明这种领域结构已简单到足以产生出关于复杂性的定理,或是不以经验的方式向我们说明,大量现实世界的问题是如何与我们用以解决它们的符号系统的能力发生关系的。这种情况也许会改变,但是在它改变之前,我们必须依赖于经验探究,通过使用我们知道如何建立的最佳问题求解程序,将经验探究作为关于问题难点的大小和特征知识的主要源泉。即使在高结构化的领域像线性规划中,理论在加强启发法——最强有力的解题算法的基础——方面所起的作用,也远比在提供对复杂性的深层分析方面所起的作用大得多。

毋需大量搜索的智能

我们对智能进行分析时,将智能与抽取和使用有关问题空间结构的信息的能力等同起来,目的是使一个问题的解能够尽可能快和尽可能直接地生成。那么,为改善符号系统解题能力而建立的新方向,也可以与抽取和使用信息的新方

法等同起来。至少有三种这样的方法得到确认。

信息的非局部用法 首先,已有好几位研究者注意到,在树搜索过程中收集的信息通常仅仅以**局部方式**使用,这样有助于在生成这一信息的特定节点上作决策。关于国际象棋阵势的信息,如果是由对后持续的次级树的动态分析获得的,通常只用于该阵势的评价,而不用于评价别的可能含有许多相同特点的阵势。所以一些完全相同的事实不得不在这一搜索树的不同节点上重复去发现。简单地把信息从它出现的语境中抽取出来,作为一般形式来使用,并不能解决这一问题,因为该信息也可能只在一个有限的语境范围中有效。近年来,在把信息从它的原始语境传递到其他恰当语境这方面,已做了一些开拓性的工作。尽管评价这一思想的力度,甚或确切说明如何实现它,还为时过早,这一思想的前景却不可小看。伯利纳(Berliner 1975)所沿袭的重要研究路线,就是使用因果分析确定出一条特殊信息的有效范围。这样,如果可以从一个国际象棋阵势的破绽追溯到造成它的那一走法,那么可以预料,同样的走法也会在别的阵势中留下同样的破绽。

HEARSAY 口语理解系统采用了另一种使信息在全局有效的方法。该系统寻求的是通过在多个不同层次上实行并行搜索来识别语音串,这些层次有:音位的、词汇的、句法的和语义的。每个这样的搜索都提出一些假设,并对其作出评价,这时它把它获得的信息提供给一块公共的"黑板",所有资源都可以阅读这块黑板。例如,这一共享的信息可以用来消去假设,甚至是整个类别的假设,否则的话,这些假设还得由一个过程来搜索。这样,就增加了我们非局部地使用树搜索信息的能力,也为提高问题求解系统的智能提供了保证。

语义识别系统　第二种提高智能的有效方式是向符号系统就其所涉及的任务域提供大量语义信息。例如，对国际象棋大师技巧所作的经验研究表明，大师的技巧主要来源于所存储的信息，这些信息有的能使他识别出棋盘上大量的特定特点和特点的模式，有的利用这种识别，提出与所识别特点相适合的行动。当然，国际象棋程序几乎从一开始就吸收了这种一般性思想。新的东西则是在数量上对这种模式的了解，以及了解可能是大师级表演所必须存储的那些联想信息，这种信息大约有 5 万个。

识别之所以可能取代搜索，是因为特殊的模式，特别是稀有模式，可能包含大量信息，只要它是与问题空间结构密切联系的。当结构"不规则"，并且不服从于简单的数学描述时，大量相关模式的知识就有可能成为智能行为的关键。任一特殊任务域的情况是否如此还是一个问题，要解决它，用经验调查比用理论更容易。我们关于符号系统的经验虽然被赋予丰富的语义信息，并被赋予很多存取这些信息的模式识别能力，然而这种经验仍然是极为有限的。

以上讨论特别谈及了与识别系统有联系的语义信息。当然，在语义信息加工和语义存储的组织方面还有整整一个 AI 研究的大领域，这不在本文讨论的论题范围之中。

选择恰当的表述方式　第三条探索路线涉及是否能够通过选择恰当的问题空间来减少搜索或避免搜索。有一个有代表性的例子可生动地说明这种可能性，即残缺国际象棋棋盘问题。标准的国际象棋棋盘有 64 个方格，每一个 1 乘 2 的长方形薄片恰好覆盖两格，用 32 个薄片正好能将整个棋盘覆盖。现在，假定将棋盘对角线上相对两角的方格割去，剩下总

共 62 个方格,这个残缺的棋盘能够正好用 31 个薄片覆盖吗?以(一丝不苟的)极大耐心,把各种可能的排列形式都试一番,可以证明这种覆盖是不能实现的。对于耐心较差而智慧较高的人来说,可采用另一种方法,由观察了解到棋盘对角线上相对两角方格的颜色是相同的。所以在残缺的棋盘上,一种颜色的方格比另一种颜色的少两个。但是每个薄片所覆盖的都是一种颜色和另一种颜色的方格各一个,任何一组薄片所覆盖的两种颜色的方格数目必然是相等的,所以这个问题的解不存在。一个符号系统在解决该问题时,怎样才能发现这种简明的归纳论证,用以替代在所有可能的覆盖方式中进行搜索的徒劳尝试呢?我们将给找到解的系统在智能上判高分。

然而,在提出这些问题时,我们也许并不是在避开搜索过程。我们只是把搜索从可能问题解答空间移置到了可能表述空间。无论如何,在问题求解研究领域中,从一种表述方式转移到另一种表述方式,以及发现和评价这些表述方式的整个过程,基本上是一块尚未开发的土地。支配表述方式的定性结构定律还有待于发现。几乎可以肯定,在未来的 10 年中,搜索它们的工作将受到极大的关注。

结　论

以上是我们就符号系统和智能所作的论述。从柏拉图的《美诺篇》到今天,走过了一条漫长的道路,然而或许令人鼓舞的是,在这条路途上的大部分进展是进入 20 世纪后取得的,而其中很大的部分又是在 20 世纪中叶以后取得的。思维在被现代形式逻辑解释为形式标记的处理之前,一直是不可

捉摸和难以名状的东西。在计算机教我们知道符号如何能由机器加工之前,它看来主要仍是存在于柏拉图的理念王国里,或是在同样模糊不清的人类心灵空间里。本世纪中叶,在这些发展处于转折点上的时候,图灵作出了伟大的贡献——将现代逻辑引入了计算机。

物理符号系统 对逻辑和计算机的研究已向我们揭示出:智能存在于物理符号系统之中。这是计算机科学最基本的定性结构定律。

符号系统是许多模式和过程的集合体,而后者能够产生前者,并破坏和修正前者。模式最重要的特性是,它们能指称对象、过程或其他模式,而当它们指称过程时,它们就能得到解释。解释的意思就是完成被指称的过程。在我们所了解的符号系统中,有两类的意义最为重大,即人类和计算机。

正如前面指出的那样,我们目前对符号系统的理解是经历了一系列阶段才增长起来的。形式逻辑使我们熟悉了符号,把符号看作思维的原材料,以句法方式来处理它们,同时也使我们熟悉了根据仔细定义的形式过程来处理符号的观念。图灵机真正以机器的方式完成了对符号的句法加工,同时也证实了严格定义的符号系统的潜在普适性。计算机的存储程序思想再次证实了已经隐含在图灵机中的符号的可解释性。表处理把符号的指称能力推到最前面,同时它定义符号加工的方式可使其独立于基础层次物质机器固定结构。到1956年时,所有这些思想都已完成,同时也获得了实现它们的硬件。对于符号系统的智能——这个人工智能的主题——的研究可以开始了。

启发式搜索 AI的第二个定性结构定律是,符号系统是

通过生成潜在可能的解,并对其进行检验,也就是通过搜索的方式,来解题的。建立符号表达式,对它们进行一系列修正,直到它们满足解的条件,这就是寻找解的一般方式。因而符号系统是通过搜索来解题的。由于它们的资源有限,这一搜索不可能立即全部完成,而必须依次进行。它留给我们的或者是从起点到目标的一条单一路线,或者是在必须修正和退回的情况下由所有这些路线组成一整棵搜索树。

符号系统处于完全混沌的环境时,不可能表现出智能。从一个问题域中抽取出信息,并运用这些信息指导搜索,从而避开错误方向和迂回曲折的分叉,符号系统是通过这种方式来行使智能的。若要使该方法有效,问题域必须包含信息,亦即包含某种程度的秩序和结构。《美诺篇》中的悖论是因注意到信息可以记住而解决的,然而新信息也有可能从符号所指称的区域中抽取出来。在这两种情况下,信息的最终源泉都是任务域。

经验基础 人工智能研究关心的是,符号系统为了具有智能行为必须怎样组织。20 年来,这一领域的工作已积累了大量知识,足以写成许多部书(也已经这样做了),其中大部分知识是以相当具体的经验形式出现的,针对的是特定任务域中特定类别符号系统的行为。然而在这种经验之中,也出现了某些一般性方式,它们超越了任务域和系统,反映出智能及其实现方法的一般特征。

我们在本文中尝试地阐述这些一般特征中的一部分。这些一般特征主要是性质上的,而不是数学上的。它们更多地带有地质学或进化论生物学的味道,而不是理论物理学的味道。它们强有力的作用,足以使我们今天为范围相当大的任

务域设计和建立中等智能的系统,同时对在许多情况下人类智能是如何工作的取得相当深入的理解。

下一步做什么? 在以上论述中,我们提到了一些已经解决的问题和一些有待解决的问题,这两方面的问题都非常之多。我们看到,在过去的四分之一世纪里,围绕这一领域进行开拓性研究的热情丝毫没有减弱。在下一个四分之一世纪里,前进的速度将取决于两种资源的限制。一是可投入使用的计算机功能的大小,二是有才干的年轻计算机科学家中被吸引到他们所能应付的最具挑战性的这一研究领域中来的人数,这后者也许是更为重要的。

A·M·图灵在他著名的文章《计算机器与智能》中以这样的话总结道:“我们的目光所及,只能在不远的前方,但是可以看到,那里有大量需要去做的工作。”

1950年时图灵认为需要去做的工作其中有不少已经完成了,但是我们的工作日程还像以往那样饱满。也许我们对他上面的简单论述理解得过于复杂了,不过我们还是倾向于认为,图灵在这里认识到了所有计算机科学家在直觉上都知道的基本真理。对于所有的物理符号系统来说,正如我们不得不对问题的环境作串行搜索一样,至关重要的问题永远是:下一步应该做什么?

参考书目

Berliner, H. (1975). 'Chess as Problem Solving: The Development of a Tactics Analayzer.' Unpublished Ph.D. thesis, Carnegie-Mellon University.

McCarthy, J. (1960). 'Recursive Functions of Symbolic Expressions and their Computation by Machine.' *Commun. ACM* 3 (Apr.): 184–95.
McCulloch, W. S. (1961). 'What is a Number, that a Man may know it, and a Man that he may know a Number?' *General Semantics Bulletin* nos. 26–7: 7–18. Repr. in W. S. McCulloch, *Embodiments of Mind*, pp. 1–18. Cambridge, Mass.: MIT Press.
Nilsson, N. J. (1971). *Problem-Solving Methods in Artificial Intelligence*. New York: McGraw-Hill.



人工智能之我见

D·C·玛尔*

人工智能是对复杂信息处理问题的研究，这些问题常常植根于生物信息处理的某个方面。该学科的目标是确定值得研究、并有可能解决的信息处理问题，然后将它们解决。

信息处理问题的解答自然地分为两部分。第一部分是对特殊计算的基础性质的表征，并对它在物理世界中的基础作出理解。这一部分工作可以看作对要计算**什么**和**为什么**计算所作的抽象的系统阐述，我把它称为计算"理论"。第二部分由实现计算的特殊算法构成，所以它说明了**怎样**做的问题。算法的选择通常视运行这一过程的硬件而定，而同一计算可由多种算法来实现。另一方面，计算理论只取决于以它为解的问题的性质。贾丁和西布森(Jardine and Sibson 1971)解构簇分析主题时，用的正是这种方法，他们的术语"方法"就是指我所说的计算理论。

为了弄清这种区别，我们看一看富里埃分析的例子。富里埃变换的(计算)理论为大家所熟知，它的表达与计算它的具体方式无关。然而实现富里埃变换的算法有好几种，如快速富里埃变换(Cooley and Tukey 1965)是一种串行算法，并行

"空间"算法是以相干光学机制为基础的。所有这些算法完成的都是同一计算，选用哪一种算法，则取决于所使用的硬件。顺便说一句，我们也注意到了串行和并行的区别是处在算法层次上的，而不是计算的深层特性。

因此严格地讲，一个人工智能结果是由以下方面组成：分离出特殊的信息加工问题，系统阐述用于该问题的计算理论，构造实现这一理论的算法，以及通过实践证明算法是成功的。对于一个特殊问题，算法理论一经建立，就再也不必重复做它，这是很重要的一点，也是可能取得进步的原因，在这方面，AI 结果的表现与数学的或任何硬自然科学的结果有其相似之处。在确定一个问题的计算理论是否已得到恰当的系统阐述时，必须采用某种判断方式。"吃掉对方的王"这一陈述规定了国际象棋的目标，但是这很难说是国际象棋计算问题的恰当表征。[①] 这里所需要的那种判断，看来与确定一个数学成果是否可作为一个新的实质性定理的判断十分类似，同时，如果未对这种判断的基础作出说明，我并不感到有什么不妥。[②]

有关由什么构成 AI 结果的这一看法，可能是大多数科学家都可以接受的。乔姆斯基（Chomsky 1965）有关英语句法的

---

* D·C·玛尔，"人工智能之我见"，引自《人工智能》9(1977)：第 37—48 页。Elsevier 科学出版社允许重印。

戴维·C·玛尔（David C. Marr），麻省理工学院视觉研究实验室主任。

① 一个在原理上能够解决国际象棋问题的计算理论，就是穷尽式搜索。然而真正有意义的做法是系统阐述人类用于下棋的计算步骤。可以假定，人们希望得到一个具有相当普遍的应用价值的计算理论，并辅之以该理论恰好可应用于某类国际象棋赛的证明，以及我们下这类棋的证据。

② 在已知的计算理论没有得到新的实质性说明的情况下，也有可能再次发明实现这一理论的新的算法，如威诺格拉德（Winograd 1976）的特快速富里埃变换，就没有对富里埃分析的性质作出任何新的说明。

"语言能力"理论的观点,恰恰就是我所说的这一问题的计算理论。它们都具有不涉及繁冗的算法细节的性质,然而在表现出语言能力(即实现计算)时,必须有算法的运行。这并不是说发明合适的算法是件容易的事,而是说,在我们可能发明算法之前,必须确切知道算法被假定用来做什么,而这一信息是通过计算理论来获取的。一个问题以这种方式解构之后,我将称它具备了**1型**理论。

美中不足的是,虽然许多生物信息加工问题具有1型理论,但是它们之所以具有这种理论的原因,却一点也不清楚。如果出现这种情况:在解决一个问题时,有众多过程同时行动,**而这些过程的相互作用就是它本身的最简单描述**,那么我称这种情况是**2型**理论。[①] 预见蛋白质怎样折叠的问题,是2型理论很有希望的候选者。当一个大肽链在媒质中震动摇摆时,有大量的因素对其施加影响。在每一瞬间,只有少数几个可能的相互作用是重要的,但是它们的重要作用却具有决定性的意义。若要建构一个简化理论,必须忽略一些相互作用;但是如果在进行折叠的某个阶段上,大多数相互作用都是关键性的,就证明简化理论是不恰当的。值得一提的是,当前最有希望的蛋白质折叠研究是以蛮力方式进行的研究,它建立了一个相当细致的氨基酸模型,有与其序列相关联的几何形状,有与周围液体的疏水相互作用,以及随机热扰动,等等,然后让这一整套过程运行,直到获得一个稳定的构形为止

---

① 这里要强调的是,在物理学中常常有一些自然模化形式(例如,在常规条件下,电的相互作用与重力的相互作用无关),但是有些过程同时包含多种模化形式,它们大致上同等重要,如蛋白质折叠。这样,1型-2型的区分就不是一个单纯的二分法,在它们之间有着一系列的可能性。

(Levitt and Warshel 1975)。

AI的根本性困难在于，一个问题是否具有1型理论，是永远不能确切肯定的。如果找到了1型理论，那当然很好；但是如果没有找到，并不意味着它不存在。到目前为止，大多数的AI程序都相当于2型理论，而采用2型理论的危险是，它们可能将一些最终为该问题的正确1型解构提供钥匙的关键性决策埋藏在那些每当设计具体程序时不可避免的成堆的小型管理决策之下。这个现象使得AI研究难于深入进行，也难于判断。如果我们证明，一个给定的信息加工问题是由一个特殊的、界定清晰的计算理论解决的，那么结果就能得到保证。反之，如果解决问题时出现的是一组冗长繁琐的过程，我们就不能完全肯定对于一个或多个相关问题不再有简单的基础计算理论存在了，因为这理论的系统阐述显得有点朦朦胧胧。对于任何一个2型理论的候选者来说，程序的性能显得重要得多。因为它唯一能够表现出的优点可能就是行之有效，所以只有它做到这一点，它才是有价值的。常常出现这种情况：一项AI研究得出的结果是一个没有多少理论含量的大程序，它对该问题的处理得出的是2型结果，但是这个程序或者性能太差，不能给人留下什么印象，或者(更糟的是)甚至无法实现。对这种研究项目的判断只能是很粗陋的，因为它们的长远作用几乎看不到。

这样我们就清楚了，AI在从事研究信息处理问题时，易于出现两种类型的求解方式。其一是传统意义上的规范的基础理论。视觉研究中有一些这样的例子：霍恩(Horn 1975)的描影构形法，表述图像密度变化和局部形状的基本框图的观念(Marr 1976)，厄尔曼(Ullman 1976)的测定光源法，宾福德

(Binford 1971)的广义柱体表述法,在此基础上的玛尔和西原(Marr and Nishihara 1978)的三维结构内部表述和处理,近期的立体视觉理论(Marr 1974; Marr and Poggio 1976)[①],以及波焦和赖卡特(Poggio and Reichardt 1976)的家蝇视觉定向行为分析。这些结果的一个特征是,在对智能功能的全面探讨中,它们往往处在较低层次上,这个层次常常是那些意在研究“更高级、更核心”的智能问题的人所不屑一顾的。对于这种批评,我们的回答是,低层次问题表现的也许的确是较容易的类型,但这正是要首先研究它们的原因。只有在解决了更多的这类问题之后,我们才会对研究较深层问题时出现的问题有更清楚的了解。

但是即使这样一些比较明显的 1 型理论,其中也包含着 2 型理论。例如,玛尔和西原的三维表述理论提出,深层基本结构是建立在一个可以看作直线型的分布式的、以对象为中心的坐标系之上的,同时这一表述方式显然在图像分析过程中得到处理。除非也能证明这种描述可以根据图像来计算,并能按照所要求的方式来处理,否则这种理论就只不过是推测而已。要这样证明,会牵涉到数个中间理论,其中有些可望最终成为 1 型状态,但另一些只能勉强地看作是 2 型状态。例如,根据对象在图像中形成的轮廓线确定恰当的局部坐标系的问题,其中有一部分目前就存在着 1 型理论(Marr 1977)。但是,对一些在基本框图上运作、以帮助图形从背景中分离出来的基本分组过程,为它们推导 1 型理论,

① 联合计算或松弛标定的概念(Zucker 1976)是算法层次上的概念。它提出了一个实现特定计算的方法,而不是着重说明应当实现什么问题,而这一点看来是关于视觉的真正争议点,其重要性不低于其他方面。

也许是不可能的。图形背景“问题”可能不是单一的问题，而是几个子问题的混合，它们结合起来实现了图形的分离，就像不同的分子相互作用而结合从而引起蛋白质折叠一样。事实上，没有理由认为图形背景问题的解答应当从单一的基础理论中得出。其原因是，它必须包含对许多有关图像事实的过程性表述，而这些图像最终还是经由物理世界中的事物所具有的聚合性和连续性演化推导而成的。这中间包含着多种知识和不同的技术，我们只好把它们一一拣出。随着每一点的积累，整体性能就得到改善，同时所能处理的图像的复杂性也在提高。

我们已经看到，如果一个问题实际上具备的是1型理论，那么寻求它的2型理论就会是危险的。这种危险最突出地表现在过早地强行进入高层次问题，因为这时作为它最终的1型理论基础层的那些概念没有或几乎没有形成，结果将是全然不能对实际所涉及的问题作出正确的系统阐述。然而，意识到较低层次上存在着反面危险，也是同样重要的。例如，在目前的视觉加工理论中，基本框图观念看来是相当不错的，但是人们也许会对使它解码的分组过程的美学特征产生怀疑。有很多这样的过程，它们的细节多少有些混乱，这样就会出现一些看起来是任意的取舍(例如，在垂直方向或水平方向形成组织)。一个明显的2型理论的例子就是我们提出的：组织与视觉的辨别是建立在这些分组过程，以及在图像基本框图中应用于该信息的一阶辨别之上的(Marr 1976)。这样，与朱尔兹(Julesz 1975)规范的(I型)理论相比，它的吸引力就小一些，朱尔兹的理论认为，只有当组织的密度分布在一阶或二阶统计中存在差异时，它们才是可以辨别的。但是正如朱尔兹本

人看到的那样，有一些模式虽然带有不同的二阶统计数据，仍然是无法辨别的。事实上，我本人所做的工作也可以看作是试图精确定义出什么样的二阶统计结构特征造成了辨别力(参阅 Schatz 1977，待印)。

我们终于迫不得已放弃了朱尔兹那个简明理论的长处，但是我感到我们不应由于在这一研究水平上需要对相当杂乱无章的细节进行开发而过于沮丧。我们已经知道，对视觉信息的其他方面——运动、立体感、荧光性、颜色——进行计算时，肯定有各别的模数存在，所以也就没有理由要求它们都建立在单一理论的基础上。诚然，人们也会预期到情况的另一面，在不断前进的进化过程中，新的模数出现了，它们能够涵盖更多方面的数据，以使动物在范围更广的环境中生存下去。仅有的重要限制是，系统作为一个整体，应该粗略地模数化，这样增添起新工具来就能比较容易。

所以即使我们找不到 1 型理论——也许就不存在这样的理论，我们也不必放弃努力，特别是在较外围的感觉信息处理的阶段上，当然也不排除在接近中枢神经的地方。更重要的是，即使有 1 型理论存在，也没有理由说这一理论同包含更多中枢神经现象的理论有着密切的联系。例如，在视觉中，有的理论认为三维表述是以直线型坐标系为基础的，并且说明了怎样处理它们，这种理论与基本框图理论无关，或者因此也与从图像到表述的大多数其他中间阶段无关。这里特别要指出，假定近似的外围过程理论对较高层次的操作有什么重要作用，是尤其危险的。例如，由于朱尔兹的二阶统计学思想是如此简洁，又与大量数据相吻合，人们也许受其影响提出这样的问题：二阶相互作用的思想是否可以以某种方式作为较高

过程的核心思想。在这样做时,不应当忘记视觉组织辨别力的真正解释,在本质上可能是全然不同的,即使该理论对视觉性能作出的正确预言非常之多。

我们所以在这一点上花如此长的篇幅，是因为它影响着另一争议点——自然语言语法所具有的理论类型问题。我们假定人类语言的目的是将一个原本不是一维的数据结构转换成一维形式，以便按顺序发声的方式传递，然后在听者头脑中被重新译作某种近似的副本。根据这一观点，完全有可能并不存在转换语法试图要定义的那种类型的英语句法的 1 型理论——这个理论规定了一些类似的硬性惯例，涉及的是执行这一冗长的但却极其重要的操作的一些实用方法，而不是关于智能性质的深层原理。抽象的句法理论有可能是一种幻想，朱尔兹的二阶统计学理论近似于一套实现组织视觉过程的行为，句法理论只是在这种意义上近似于真实情况，而这一套过程归根结底就是这个理论的全部内容。换句话说，自然语言语法具有的很可能是 2 型理论，而不是 1 型理论。

即使生物信息处理问题只有 2 型理论,还是有可能从它的解答中推论出比解答本身更多的东西。这种情况的出现是因为,在实现一组过程中的某个点上,那些附加在机器上的以使机器运行的设计规定,会开始对实现方式的结构产生影响。这一观察结果为语言学家和人工智能圈内的学者进行的两种类型的研究增添了不同的视角。如果句法理论确实是 2 型的,那么有关 CNS 的任何重要的内在意义,都有可能从实现它的组成过程的方法细节中得到,这些内在意义往往只有通过实现这些过程才能得到开发。

1. 本观点意义所在

如果接受了关于 AI 研究的这一观点，就可以根据较为清楚的判据来判断它的成果。已离析出什么样的信息处理问题？解决它的规范理论已经形成了吗？如果形成了，支持它的论据的充分程度如何呢？如果规范理论尚未给出，那么支持一组过程解答的证据是什么，或者指出它不存在单一规范理论的证据是什么呢？以及，所提出的这套机制工作得顺利吗？对于像理解故事这样相当高层的问题，当前的研究往往是纯探索性的。也就是说，在这些领域中，我们的知识还相当贫乏，我们甚至还无法开始归纳出恰当的问题，更不用说解决它们了。必须看到，个人的冒险，这是人类作任何尝试时都难以避免的局面（几乎可以肯定，所有做探索的先驱者本人在寻找实用性问题方面都是不成功的），但是这是最后成功所必不可少的前奏。

AI（现在已满 16 岁了）的大部分历史是由探索性研究构成的，其中最著名的有斯莱格尔（Slagle 1963）的符号积分程序，魏岑鲍姆（Weizenbaum 1965）的 Eliza 程序，埃文斯（Evans 1968）的模拟程序，拉斐尔（Raphael 1968）的 SIR，奎连（Quillian 1968）的语义网络，和威诺格拉德（Winograd 1972）的 Shrdlu。（回顾）所有这些程序，其特点是或者太简略，构不成有价值的 1 型理论，或者虽然非常复杂，可是性能太差劲，也不能严格地作为 2 型理论看待。AI 早期出现的真正成功的 2 型理论，也许只有沃尔兹（Waltz 1975）的程序。但是我们从这些经验

中学到了很多东西——大多是反面教训(例如,关于智能有可能是怎样工作的这一问题的已知的前20个观点,要么太简单,要么是错误的),当然也包括若干正面经验。MACSYMA代数处理系统(Moses 1974)无疑是成功和有用的,它植根于一些程序,如斯莱格尔的程序。这个领域中出现的错误,不在于进行了这样一些研究——它们构成了AI发展的基本方面,而主要在于错误地判断了这些研究的价值,因为早期研究本身几乎没有归纳出任何可解的问题,这一点现在已经很清楚。这些内部判断失误的原因,部分在于这一领域早期成果受到的外部压力,但是这终究是一些政治问题,这里不打算讨论。

然而,我认为,人们在对这些判断的错误作出判断时,又错误地采取了过于苛刻的态度。它们只是必要的热情所产生的难以避免的结果,其出发点是认为这一领域具有持久的重要性,在我看来这是正确的。人类锲而不舍的所有重要事业,都是契而不舍地以基于信念、而不是基于后果的个人献身精神作为开始的。AI正是这样的例子。只有偏狭的、爱挑剔的、缺乏冒险精神的人才会用它作为反对我们的理由。

2. 当前趋势

探索性研究是重要的。这一领域中的许多人都怀有这种期望:在我们理解智能的核心思想深处,至少会有一个、也可能是若干个关于怎样组织和表述知识的重要原理,从而在某种意义上弄清了什么是有关我们智力一般性质的重要东西。乐观主义者或许在一些程序如萨斯曼和斯托尔曼(Suss-

man and Stallman 1975)的程序、玛尔和西原(Marr and Nishihara 1978)的程序中,在明斯基(Minsky 1975)就核心问题提出的全面见解中,还可能在尚克的一些工作(Schank 1973, 1975)中,看到这种原理的少许端倪,虽然我有时感到尚克没有抓住要点。尽管还有些疑云,如下一些观点看来正在兴起(它们在很大程度上归功于早期的探索性研究):

1. 有关推理、语言、记忆和感知的"组块"应该比当前心理学理论所容许的大多数情况更大一些(Minsky 1975)。它们还必须是非常灵活的——至少与玛尔和西原的直线型三维模型一样灵活,或许更加灵活。由"框架"、"端点"这些术语提出的简易机制当然过于死板了。

2. 对一个事件或一个对象的感知必须包含对它的几种不同描述的联立计算,这些描述涉及事件或对象的用途、目的或环境的一些不同方面。

3. (2)中提及的各种描述既有粗略形式,也有精细形式。在根据(1)中的要求选择恰当而全面的情节梗概时,以及正确规定那些造成这些情节梗概被选定的对象和行动所起的作用时,这些粗略描述是一个很重要的环节。

用一个例子可以更清楚地说明这些观点。如果某人读到:

(A) 一只苍蝇在玻璃窗上讨厌地嗡嗡叫。

(B) 约翰拿起报纸。

其直接推论是:约翰对这只苍蝇的态度基本上是厌恶的。假如他拿起的是电话机,推论就不会这么肯定。我们都会承认,读到这些句子时,一段"伤害昆虫"的情节梗概以某种方式展开来,它是通过苍蝇讨厌地嗡嗡叫以最粗略的方式提示的。

这一情节梗概包含一个参照物，是某个有可能把这个小虫子在玻璃上压扁的东西，这一描述适用于报纸，而不适用于电话机。我们还可以得出这样的结论：在提到（在视觉场合是“看到”）报纸时，不仅以内部方式把它描述为报纸，以及对它具有的形状和轴作某种粗略的三维描述，而且它也被描述为轻的、柔韧的扁平物体。因为句子（B）之后也可能是“坐下来阅读”，所以报纸也可能被描述为阅读材料；照此类推，还可以作为易燃物，或其他东西。既然我们一般事先不知道一个对象或行动的哪个方面是重要的，所以在相当一段时间里，已知对象会引起若干种不同的粗略内部描述。对行动来说，情况也是类似的。重要的也许是要注意到，对拍苍蝇、阅读或点火的描述不是非得与报纸联在一起不可，对报纸来说，只有与每一情节梗概中的作用相匹配的描述才是有用的。

尚克的“基本行动”的重要之处，依我看，既不是它们的数目凑巧不是很大这个事实，也不是每一幕的情节完全通过归纳为这些行动而得到表现的这一思想（我根本不相信这一点），甚至也不是它们所关联的情节梗概包含了当前情境的全部答案这一思想（这正是灵活性丧失作用的地方）。事件和对象的基本的、粗略的目录所具有的重要性，存在于这种粗略描述在最终接近并构造或许是精心剪裁的特定情节梗概时所起的作用，其方式很可能是这样的：存储于基本模型中的图像和信息经过适当的相互作用之后，玛尔和西原理论中的一般三维动物模型最后可能变成一只十分特殊的柴郡猫。在句（A）之后，原先只不过是对无辜苍蝇的厌恶态度，随着报纸信息的加入，就变成了一个拍苍蝇的特定场景。

玛尔和西原把为报纸提供多重描述的问题称为报纸的

“参照窗口问题”。怎样做得最好，什么样的描述能够适合于不同词汇或不同感知对象，目前尚无确切了解。这些见解是探索性研究的结果，它们引出的问题尚需作准确的系统阐述，还谈不上得出令人满意的解答，但是现在可以肯定，某些这种类型的问题的确是存在的，也是重要的，而且看来最终很可能会出现一个有关这些问题的相当有影响的理论。

3. 模仿还是探究

最后，我打算作出进一步的区分，这对于选择研究问题，或是判断已完成工作的价值看来都是重要的。这个问题就是，研究工作，特别是对自然语言理解、问题求解或记忆结构的研究，很容易蜕化成为编写程序，这种程序只不过是一种没有启迪作用的对人类行为方式的某个小方面的模仿而已。魏岑鲍姆(Weizenbaum 1976)现在认为他的程序 Eliza 就属于这一范畴，而我也的确看不出反对他的理由。基于同样的立场我也批评了纽厄尔和西蒙在产生式系统方面的工作，以及诺尔曼和鲁梅哈特(Norman and Rumelhart 1974)在长时记忆方面的某些工作，这些批评引起较多的争论。

其原因如下。如果我们认为信息处理研究的目的是系统阐述和理解特定的信息处理问题，那么处于核心地位的正是那些问题的结构，而不是实现这些问题所经历的机制。这样，需做的第一件事就是找到那些我们有把握解决的问题，弄清怎样去解决它们，并根据这种理解来查看我们的行为方式。这些问题最丰富的源泉就是我们熟练自如地(因而是无意识

地）完成的操作，因为如果没有健全的基础方法，很难弄清可靠性怎样能得到保证。另一方面，对问题求解的研究已趋于集中在那些我们在智能上理解得很好、但做起来却很差的问题上，像心算和折算，或是集中在像几何定理证明问题或棋类游戏上面，在这些问题中，人的技巧似乎建立在庞大的知识和技能基础之上。我认为，这是现在还不去研究人类怎样完成这些任务的特别有说服力的理由。我不怀疑，在做心算时，我们正在熟练地做某件事，但它不是算术，而且在我看来，对这件事是什么，我们连哪怕一个方面也远没有理解。所以我们应首先专注于比较简单的问题，这是我们有希望取得真正进步的地方。

如果无视这一限制，最终得到的将是一些不太像样的机制，它们唯一的长处是，我们做不到的事情它们也做不到。依我看，产生式系统的特点正好与这种情况相符。即使按它们本身的说法，把它们看作机制，也还是有许多不尽如人意的地方。作为编程语言，它们设计得太差劲，难以使用，同时我也无法相信，人类大脑会在如此低等水平上用如此拙劣的实现方式担负起决策的任务。

供学生做问题求解使用的产生式系统和供视觉神经生理学家使用的富里埃分析，这两者之间也许可以作一类比。以图像的空间频率表述方式做简单操作，可以模拟不少有趣的视觉现象，它们看上去就像是由我们的视觉系统表现出来的。这些现象包括检测重复性、某些视错觉、离散的线性叠加信道的观念、整体形状与精细的局部细节的分离、大小不变性的简单表达。在图像分析中，空间频域之所以被忽略，是因为它对视觉的主要工作——根据密度分布形成“存在的是什么”的描

述，实际上不起作用。这个过程有可能怎样完成，这种直觉是视觉生理学家们所缺乏的，而这一点又是如此重要。产生式系统展示出不少有价值的思想——取消直接的子程序调用、采用黑板交流渠道，以及某个短时记忆观念。但是正因为产生式系统表现出这些副作用（就像富里埃分析"表现"某些视错觉），所以并不意味着它们与实际上正在进行的事情没有什么关系。例如，据我个人推测，短时记忆能像一个存储记录器一样起作用，这个事实在它的诸种功能中可能是最不重要的。我估计，有若干"智能本能反射"作用于短时记忆拥有的条目，对此我们还一无所知，而这些本能反射最终会被看作对它来说是至关重要的事情，因为它们起着主导作用，如为一个条目打开参照窗口。依我之见，从与产生系统有密切关系的角度来研究人类的表现是白费时间，因为这等于是在研究一个机制，而不是在研究一个问题，因而不可能导致 1 型结果。这种研究试图透彻了解的那些机制将会通过研究问题而被揭示出来，正如视觉研究不断进步是由于它攻克的是视觉**问题**，而不是视神经机制。

对同一批评意见的反思可见之于诺尔曼和鲁梅哈特的著作，他们研究的是长时记忆中可能存在的信息组织方式。我们再次看到，危险在于没有对有关的明显的信息处理问题提问。相反，所建议的提问和作答根据的是机制——这时它被称为"主动结构网络"，它是如此简单和笼统，以致缺乏理论实质。他们或许会说，如此这般的"联系"似乎是存在的，但是他们无法说明，联系是由什么构成的，也无法说明因为要解决问题 X（是我们所能解决的），需要一种用如此这般的方式组织的记忆，所以该联系必须如此；如果谁具有这种联系，某些明

显的“联系”就会作为副产品而出现。在发现一些需要作出解释的事实方面，实验心理学的确是卓有成效的，这些事实包括关于长时记忆的事实，同时（例如）谢泼德（Shepard 1975）、罗希（Rosch，付印中）和沃林顿（Warrington 1975）的工作，在我看来都是这方面非常成功的例子。但是和实验神经生理学一样，如果信息加工研究还没有确认和解决恰当的问题 X，实验心理学就不可能对这些事实作出解释。[①] 依我之见，找出这样的问题 X，并解决它们，就是 AI 应该尝试去做的事情。[②]

参考书目

Binford, T. O (1971). 'Visual Perception by Computer.' IEEE Conf. Systems and Control, Miami.

Chomsky, A. N. (1965). *Aspects of the Theory of Syntax*. Cambridge, Mass.: MIT Press..

Cooley, J. M., and Tukey, J. W. (1965). 'An Algorithm for the Machine Computation of Complex Fourier Series.' *Math. Comp.* 19: 297-301.

Evans, T. (1968). 'A Program for the Solution of Geometric-Analogy Intelligence Test Questions.' In M. Minsky (ed.), *Semantic Information Processing*, pp. 271–353. Cambridge, Mass.: MIT Press.

Horn, B. K. P. (1975). 'Obtaining Shape from Shading Information.' In P. H. Winston (ed.), *The Psychology of Computer Vision*, pp. 115–55. New York: McGraw-Hill.

Jardine, N., and Sibson, R. (1971). *Mathematical Taxonomy*. New York: Wiley.

Julesz, B. (1975). 'Experiments in the Visual Perception of Texture.' *Scientific American* 232: 34–43.

---

① 就目前的技艺状况来说，最明智的作法看来是集中研究那些可能有 1 型解的问题，而不是去研究那些几乎可以肯定是 2 型的问题。

② 我这里提出的纯粹是个人的观点，对此我负完全责任，但是它们可能有什么价值的话，应部分归功于我同 D·麦克德莫特的多次交谈。本文提及的工作是在麻省理工学院 AI 实验室完成的。国防部高级研究项目署为该实验室的 AI 研究提供了资助，属于海军科研局 N00014－75－C－0643 号合同。

Levitt, M., and Warshel, A. (1975). 'Computer Simulation of Protein Folding.' *Nature* 253: 694–8.

Marr, D. (1974). 'A Note on the Computation of Binocular Disparity in a Symbolic, Low-Level Visual Processor.' MIT AI Lab. Memo 327.

—— (1976). 'Early Processing of Visual Information' *Phil. Trans. Roy. Soc. B* 275: 483–524.

—— [1977]. 'Analysis of Occluding Contour.' [*Proc. Roy. Soc. London, B* 197: 441–75.]

—— and Nishihara, H. K. [1978]. 'Representation and Recognition of the Spatial Organization of Three Dimensional Shapes.' [*Proc. Roy. Soc. London, B* 200: 269–94.]

—— and Poggio, T. (1976). 'Cooperative Computation of Stereo Disparity.' *Science* 194: 283–7.

Minsky, M. (1975). 'A Framework for Representing Knowledge.' In P. H. Winston (ed.), *The Psychology of Computer Vision*, pp. 211–77. New York: McGraw-Hill.

Moses, J. (1974), 'MACSYMA—The Fifth Year.' *SIGSAM Bull., ACM* 8: 105–10.

Norman, D. A., and Rumelhart, D. E. (1974). 'The Active Structural Network.' In D. A. Norman and D. E. Rumelhart (eds.), *Explorations in Cognition*, pp. 35–64. San Francisco: W. H. Freeman and Co.

Poggio, T., and Reichardt, W. (1976). 'Visual Control of the Orientation Behavior of the Fly: Towards the Underlying Neural Interactions.' *Quarterly Reviews of Biophysics* 9: 377–438.

Quillian, M. R. (1968). 'Semantic Memory.' In M. Minsky (ed.), *Semantic Information Processing*, pp. 227–70. Cambridge, Mass.: MIT Press.

Raphael, B. (1968). 'SIR: Semantic Information Retrieval.' In M. Minsky (ed.), *Semantic Information Processing*, pp. 33–145. Cambridge, Mass.: MIT Press.

Rosch, E. (in press). 'Classification of Real-World Objects: Origins and Representations in Cognition.' *Bulletin de psychologie.*

Schank, R. C. (1973). 'Identification of Conceptualizations Underlying Natural Language.' In R. C. Schank and K. M. Colby (eds.), *Computer Models of Thought and Language*. San Francisco: W. H. Freeman.

—— (1975). *Conceptual Information Processing.* Amsterdam: North-Holland.

Schatz, B. R. (1977). 'The Computation of Immediate Texture Discrimination.' MIT AI Lab. Memo 426.

Shepard, R. N. (1975). 'Form, Formation, and Transformation of Internal Representations.' In R. Solso (ed.), *Information Processing and Cognition: The Loyola Symposium*, pp. 87–122. Hillsdale, NJ: Erlbaum.

Slagle, J. R. (1963). 'A Heuristic Program that Solves Symbolic Integration Problems in Freshman Calculus.' In E. A. Feigenbaum and J. Feldman (eds.), *Computers and Thought*, pp. 191–203. New York: McGraw-Hill.

Sussman, G. J., and Stallman, R. M. (1975). 'Heuristic Techniques in Computer-Aided Circuit Analysis.' *IEEE Transactions on Circuits and Systems*, CAS-22: 857–65.

Ullman, S. (1976). 'On Visual Detection of Light Sources.' *Biol. Cybernetics* 21: 205–12.

Waltz, D. L. (1975). 'Understanding Line Drawings of Scenes with Shadows.' In P. H. Winston (ed.), *The Psychology of Computer Vision*, pp. 19–91. New York: McGraw-Hill.

Warrington, E. K. (1975). 'The Selective Impairment of Semantic Memory.' *Q. J. Exp. Psychol.* 27: 635–57.

Weizenbaum, J. (1965). 'ELIZA—A Computer Program for the Study of Natural Language Communication between Man and Machine.' *Commun. ACM* 9: 36–45.

Weizenbaum, J. (1976). *Computer Thought and Human Reason*. San Francisco: W. H. Freeman.
Winograd, S. (1976). 'Computing the Discrete Fourier Transform.' *Proc. Nat. Acad. Sci.* 73: 1005–6.
Winograd, T. (1972). *Understanding Natural Language*. New York: Academic Press.
Zucker, S. W. (1976). 'Relaxation Labelling and the Reduction of Local Ambiguities.' University of Maryland Computer Science TR-451.

7 认知之轮：人工智能的框架问题

D·C·丹尼特*

从前有个机器人，制造它的人给它起名叫 R_1。它只有一个任务，就是照料自己。一天，在设计者的安排下，它得知它的备用电池——它珍贵的能源——和一个快要爆炸的定时炸弹锁在一间房子里。R_1 找到了这个房间和房门钥匙，并做出抢救电池的计划。房间里有一辆小车，电池就在这辆车上。R_1 假设，某个叫做**拉出（小车，房间）**的行动能将电池从房间里转移出来。它立即行动，果然在炸弹爆炸前将电池从房间里取出。然而不巧的是，那只炸弹也在小车上。R_1 虽然**知道**炸弹就在房间里的小车上，但是没有意识到拉小车时炸弹会随着电池一起被带出来。可怜的 R_1 在计划它的行动时遗漏了这个明显的蕴涵关系。

回到制图板前。设计者们说："其解答很明显。我们的下一个机器人，一定要造得不仅能识别它的动作中的拟议的蕴涵关系，而且也能识别这些动作附带的蕴涵关系，可通过它做计划时采用的那些描述来推演这些关系。"他们把下一个模型——机器人推演者叫做 R_1D_1。他们把 R_1D_1 放到与 R_1 失败时相同的险境中。当 R_1D_1 也产生**拉出（小车，房间）**的想法时，

就像设计的那样开始考虑这种行动过程的蕴涵关系。它刚刚推演完把小车从房间里拉出来不会改变房间墙壁的颜色，正要着手证明下一个蕴涵关系——拉出小车时会造成它的轮子转的圈数比小车轮子的多，就在这时，炸弹爆炸了。

回到制图板前。设计者们说："我们必须教它区分开相关蕴涵关系和无关蕴涵关系，还要教它忽略那些无关的。"于是他们想出一个方法，给蕴涵关系加上标记，标明它与当前任务是相关的还是无关的，并在下一个模型中采纳了这一方法，这一模型叫做机器人相关推演者，简称 R_2D_1。制造者们让 R_2D_1 去接受那个专门设计的、曾使它的前任们丧生的试验，这时他们惊奇地看到，它正坐在那间装有嘀答作响的炸弹的房子外面，一幅哈姆雷特的做派，它果断的精神本色因陷入沉思而显出病容，正像莎士比亚(以及最近的福多尔)生动描写的那样。"干点什么吧！"他们朝它喊。"我正在做，"它反驳说，"我正忙着忽略成千上万我已确定为无关的蕴涵关系。我只要发现一个无关的蕴涵关系，就把它放进那些必须忽略的关系表中去，并且……"炸弹响了。

所有这些机器人都受到**框架问题**的困扰。[①] 如果要让机

---

* D·C·丹尼特，"认知之轮：人工智能的框架问题"，引自《心灵、机器及进化：哲学研究》，(1984)，第 129—151 页，剑桥大学出版社允许重印。

D·C·丹尼特(Daniel C. Dennett)，塔夫特大学哲学教授、认知研究中心主任。

① 该问题是由 J·麦卡锡和 P·海斯在 1969 年的论文中提出的(McCarthy and Hayes 1969)。麦卡锡在 1960 年首次对出现该问题的研究课题作了详细阐述。我对 J·麦卡锡、P·海斯、B·莫尔、Z·佩里舒、J·豪格兰和 B·达尔布姆表示感谢，他们为了让我理解框架问题，花费了很多时间。他们的教导还有很多未起作用，这不是他们的过错。

我在阅读"模型变化：框架问题"时也获益良多，这篇文章尚未发表，作者是瑞典于默奥大学信息处理研究所的 L·－E·杨勒特。我期待着这篇文章的下一个稿本不久即能问世，因为它对初学者来说是非常宝贵的便览手册，同时提出了若干新颖的主题。

器人 R_2D_2 像故事里那样聪明，并具有实时(real-time)的灵巧，那么，机器人的设计者们就必须解决框架问题。起初看来这充其量是机器人学中的一个令人烦恼的技术障碍，或者不过是令从事人工智能研究的人一筹莫展的一道奇特的难题。对此，我有不同看法，我认为这是一个新的深层认识论问题——一代代的哲学家原则上能理解，但却未加注意的问题，这一问题是由 AI 的一些新方法揭示出来的，然而还远远没有解决。许多从事 AI 的人已经开始从同样的高度来看待框架问题的严重性。正如一个研究者自嘲的那样，"我们已经放弃了设计智能机器人的目标，转而去设计一杆枪，用它去摧毁别人设计的任何智能机器人！"

这里，我打算对框架问题作一个基础的、非技术性的、哲学方面的介绍，并说明它为什么如此倍受关注。我没有什么现成的解答，甚至也不能就解答可能存在于何处提出什么独到见解。我发现，仅仅说清楚框架问题是什么，又不是什么，已经相当困难。事实上，在 AI 研究圈内，也没有就其用法完全达成一致。创造这一术语的麦卡锡和海斯，用它表示一个产生于小范围的特殊的表述问题，它仅仅出现在某些处理较大范围的实时计划系统问题的策略中。另一些人把这个较大范围问题称为框架问题——"整块布丁"，如海斯称它的那样(个人通信中)，这可能不仅仅是术语上的不当。如果产生于小范围的问题的"解答"能把一个(更深层的)困难驱入大范围问题的某个另外的角落中去，我们最好把这个名称留给这个死角中的困难。我对麦卡锡和海斯感到抱歉，因为我将参加到擅用他们术语的人中去，试图介绍这个整块的布丁，并把它称为框架问题。以后在适当的时候我将尝试描述这一问题的

较小范围的形式,也可以称之为"狭义框架问题",同时我还要略为说明它与较大范围问题的关系。

不管框架问题是什么,它必定是尚未解决的(从它目前的外表看,也可能是无解的),因此,AI 思想上的反对者们,像 H·德雷福斯和 J·塞尔,竟然为这一领域编写挽词,并把框架问题作为它死亡的原因。在《计算机不能做什么》(Dreyfus 1972)一书中,德雷福斯试图证明,对心灵研究来说,AI 是一个根本错误的方法。其实,他对 AI 模型多少有些主观的种种抱怨,以及他宣称的对于 AI 模型固有局限性的种种见解,也可以看作是相当系统地在框架问题相邻领域中的巡视。德雷福斯从未明确提到过框架问题,[①] 但是这会不会就是他曾在寻找但是不很知道怎样描述的那支冒烟的手枪呢? 诚然,我认为可以说 AI 正握着一支冒烟的手枪,但是至少在"整块布丁"的外表下,这是一个人人都难以逃脱的问题,而不只是 AI 的问题,AI 就像许多神秘故事中善良的丑人一样,它应该被誉为有所发现,而不是被指控为犯罪。

人们不一定预期充满机器人的未来世界会受到框架问题的困扰。这个问题显然来自某些广为流行的、**看起来**无害的

---

① 德雷福斯提到了麦卡锡(McCarthy 1960:213—14),但是他所谈的主题是,麦卡锡忽视了**物理状态**描述与**情境**描述之间的差别,这一主题也许可以简略地概括为:房子并不是家。

同样,他还提到**其余情况相同**(ceteris paribus)的假定(在该书修订版长达 56 页的导言中),但是他只是申明信守维特根斯坦的这一思想:"任何时候,如果是根据规则来分析人类的行为,这些规则中必须总是包括**其余情况相同**的条件……"。然而,即使这种说法正确,它也遗漏了更深层的问题:对像**其余情况相同**假定这类东西的需要,困扰着鲁宾逊(笛福《鲁宾逊漂流记》中的主人公——译者注),正像它不可避免地困扰着任何一位发现自己处在受人类文化影响情境中的主人公一样。看来这个问题并不只是局限于**人文科学**(人们常常这样认为);事实上一个在(另一个?)渺无人迹的、但却不无敌意的星球上的"智能"机器人一旦筹划它的生活,它就面临着框架问题。

有关智能性质的假定，也来自标榜为最不尚空谈的物理主义的真理，还来自这样的信条："我们怎样思维"必定是能够解释的。（二元论者回避框架问题——但这只是由于二元论揭开了那张盖在所有棘手的"怎样做"问题上的神秘而朦胧的面纱。正如我们将要看到的，一旦人们认真回答某些"怎样做"问题时，框架问题就会出现。二元论者想通过框架问题为自己辩解，这是办不到的。）

智能人的最为核心的特征——如果不说它是限定性特征的话——就是它能够"先看而后行"。更好一些则是，它能够先思而后行。智能（至少部分）是一个恰当使用你的认识的问题，但是这有什么用处呢？它可以提高预见下一步发生什么的准确性，可以作出计划，可以对行动过程进行周密思考，可以以增加将来要用的知识为目的构造进一步的假设，这样你就能让你所作的假设代你去死，从而保全你自己（正如K·波普爵士曾经说过的那样）。愚蠢（而不同于无知）的人是这样的人，他为了看燃料箱而擦亮火柴[①]，他锯断了自己坐在其上的树枝，他把钥匙锁在汽车里，又花了一个小时苦思冥想，究竟怎样使他的家人从汽车里出来。

但是在三思而行时，**我们是怎样做的呢**？答案似乎很明显：智能的人是从经验中学习，然后用学到的东西来指导对未来的预期。实际上，休谟就是根据预期的习惯对此作出解释的。**但是习惯是怎样起作用的呢？**休谟信手拈来一个答案——联结主义，大意是说：观念之间的某些转换通路，在变得

---

① 这个例子来自C·切尔尼亚克在"理性和记忆结构"中关于理性的重要论述（Cherniak 1983）。

经久耐用之后就更有可能循之而行，然而既然对于这些联接的技术细节作出更详细的解释，肯定不是**休谟**的职责，所以关于如何恰当使用这种通路——而不只是把它们转变成一个由一些无法通过的替代物组成的走不出的迷宫——的问题并没有得到揭示。

其实，休谟像所有其他的哲学家和“心灵主义”心理学家一样，无法领会框架问题，因为他是在我称为纯语义的层次上，或是**现象学**层次上从事研究的。在现象学层次上，所遇到的任何条目都**有其独立意义**。这种意义也可以看作是“给定的”——但这不过表明了理论家擅自给出了他所希望的所有意义。照此办理，一个条目与下一个条目之间的语义关系看起来一般是清楚明白的，人们只要假定条目表达的内容正好与具有那些意义的条目所**应该**表达的相符即可。通过综述休谟式的关于学得一点知识的说明，我们就可以得出这一看法。

假定有两个孩子，都有天生不经请求就从罐里拿甜饼干的倾向。一个孩子被允许这样做，不受指责，而另一个每次一伸手，屁股就要挨打。结果如何？第二个孩子学会了不去拿饼干。原因何在？因为她已有拿饼干后紧接着屁股挨打的经验。这样做有什么用处呢？拿饼干的**观念**通过习惯通路变得同打屁股的观念有了联系，接着又同疼痛的观念有了联系……这样，孩子**当然**就变得克制了。何以如此？这正是有那种环境就有那个观念的结果。但是，原因何在呢？在类似的情况下，疼痛观念还应该有什么别的作用呢？它有可能让孩子用左脚尖做芭蕾舞的旋转动作，或者背诵诗歌，或者眨眼睛，或者回忆起她的5岁生日。但是若要给出疼痛观念**意指**的内容，这些结果无一不是荒谬的。的确如此；那么在已知观

念意指的东西后，**怎样**设计观念才能使它们的结果就是应有的结果呢？设计某些内部事物——姑且称其为观念——使得它的行为与它的同类相比较时好像它意指的就是**饼干**或**疼痛**，是使得那个事物具有那种意义的唯一方式。如果它不具备这些内部行为特性，它就不能意指一个事物。

这是哲学家们留给未来某个隐约存在的研究者的技术细节问题。这种分工似乎没有什么不妥，但是这样一来，学习和智能中的大多数真正的困难和深层难题就被拒之门外。这犹如哲学家们声称自己完全有把握解释舞台魔术师采用的方法，可是在请他们解释魔术师怎样变出把女郎锯为两半的戏法时，他们的解释是：这真的是非常显然的，魔术师并没有真的把她锯成两半，他只不过做得像那么回事。“但是他是怎样完成**这件事**的呢？”我们问。“这不是我们的事，”哲学家们说——而且他们中有些人还响亮地补充道：“解释必须在某处停止。”[1]

当一个人在纯现象学或语义层次上进行操作时，他从哪里得到所需的资料，他又是怎样进行理论归纳的呢？“现象学”这一术语在习惯上是同内省法联系在一起的——用它来**检验**呈现意识给的是什么，或者赋予意识的是什么。根据定义，个人的现象学只是他或她意识中的内容。虽然这是长久以来的思想观念，但是这决非是实践。例如，洛克可能认为他的“历史的、朴素的方法”是一种无偏见的个人观察法，但事实上，这在很大程度上是一种关于观念和印象为了完成那些它

---

① 注意，在这一如实写照中，哲学家们还是做了某些有益的研究；想一想人们可能因之而消除某种徒劳无益的想法：某个研究者轻率地下结论说魔术师真的把那个女郎锯作两半然后又把她神奇地合在一起。人们毕竟曾简单地得出过这样愚蠢的结论，例如许多哲学家也曾这样做过。

们"显然"做过的工作而**必须是**什么的伪装起来的先验推理。[①] 那种关于我们每个人都能观察到我们心理活动的神话,使那个认为只要我们对自己的状况仔细地加以反省,就可以在思维理论上取得重要进步的错误看法得以延续。现在我们已经在一段时间内取得进一步的认识:我们的意识所抵达的,可以说只不过是我们头脑中发生的多层信息加工系统的外层表面而已。然而,这个神话还是造成了它的牺牲品。

舞台魔术师的比喻相当贴切。一个人要弄清这个戏法是**怎样**变的,如果只是专注地坐在观众席上目不转睛地盯着看,是不大可能取得什么进展的。视线以外的东西太多了。最好还是正视事实:我们必须对后台或舞台侧翼作彻底检查,以期用有效的方式剖解这个表演;反之,则是坐在扶手椅里,对于**必须**怎样变这个戏法进行推理思考,而不管明确规定的约束条件是什么。因此,框架问题颇像那种常见的、但却不解决问题的"发现"——就扶手椅式思考所能作出的判断而论,我们刚刚观察到的这个戏法是完全不可能的。

这儿有一个变戏法的例子——制作午夜快餐。我怎样能为自己搞一份午夜快餐呢? 怎样做比较简单? 我猜想冰箱里有一些剩下的火鸡片和蛋黄酱,面包箱里有面包,而且冰箱里还有一瓶啤酒。我意识到可以把这些东西凑到一起,于是就炮制了一个十分简易的计划:我只要走过去,察看一下冰箱,拿出所需的材料,为自己做一份三明治,就着啤酒吞下去。我还需要刀子、盘子和啤酒杯。我立即将这一计划付诸实行,它成功了! 太棒了。

① 见我的文章(Dennett 1982a),这是对古德曼文章(Goodman 1982)的评论。

不了解大量情况，我当然做不到这一点——这些情况涉及面包，摊开蛋黄酱，打开冰箱，还有磨擦力和惯性——它们在我端着盘子去我座椅旁的桌子时使火鸡能夹在面包片内，面包能留在盘子里。我也需要了解，怎样使啤酒从瓶里出来进入玻璃杯。[1] 幸运的是，多亏我有以前积累起来的关于现实世界的经验，我可以用所有这些世间的知识把自己武装起来。当然，我所需的有些知识**也许**是天生的。例如，我必须知道这样一件小事：啤酒倒进杯子之后，它就不在瓶子里了。还有，如果我正用左手拿着蛋黄酱罐，我就不能再用它来拿刀子摊开蛋黄酱。这些情况也许同某些更基本的事情有直接的蕴涵关系，是其具体表现，这些更基本的事情，我实际上是**生而知之**的，比如这样的事实：如果某物在这一位置上，就不可能也在另一不同位置上；两个物体不可能同时处在同一地方；状态的变化是行动引起的。仅凭想象很难得知人们怎样能从经验中了解到这些事实。

在我们行动和做计划时，像这样最平常的事实逃过了我们的注意，同时我们也不必奇怪以现象学方式，**而不是内省方式**作思考的哲学家们必然会忽略它们。但是如果无视内省，完全以"异质现象学"[2] 的方式思考这一任务提出的纯信息要求：每一个能够完成这一任务的实体**必须**了解哪些情况，那么这些平常的知识就会引起我们的注意。我们不难使自己感到满意的是，执行者如果不**以某种方式**受惠于这样的信息（瓶

---

① 这种物理学知识，不是在学校里学到的，而是在童床上学到的。见 Hayes 1978, 1979。

② 关于对异质现象学的详细讨论，见 Dennett 1978, Ch. 10，"心理意象的两种方式"，以及 Dennett 1982b。也见 Dennett 1982c。

中的啤酒不会在杯中,等等),就不能完成如此简单的任务。AI使人们成为这一改进方式下的现象主义者,这是它在方法论上的主要优点之一。作为异质现象主义者,其推理方式是:执行者为了用各种不同的方式完成任务而必须**无意识或有意识地**“知道”什么或断定什么。

AI迫使平常的信息浮现出来,其原因是AI面对的任务是从零开始的:通过编程去模拟执行者的计算机(或是机器人的脑,如果我们确实准备在真实的、非模拟的世界中进行操作的话),对“关于现实世界”的事最初是一无所知的。计算机是传统说法中的**白板**,它需要的每一条目都必须以某种方式印上去,或者在开始时由编程员来做,或者通过系统的后续“学习”来做。

今天我们都会同意,一个出生时像**白板**一样面对世界的实体,根本不可能学习,但是在什么是天生的、什么是在成熟中形成的以及什么是真正学习所得的之间划分出界线,比起人们或许已经产生的看法,在理论上就不那么重要了。虽然有些信息必须是天生的,但是几乎没有什么特殊条目必须是:体察到或许是**肯定前件的假言推理**,还有排中律,以及对因果性的某种观念。同时虽然有些事情我们必须学习才能知道,如感恩节在星期四到来,冰箱可以使食品保鲜,但是另外许多“完全经验”的事情,原则上是可以生而知之的,如微笑意味着愉快,没有悬挂好、没有支撑的东西会掉落。(事实上,有某种证据表明,人有一种容易感知以重力加速度降落物体的天生倾向。)①

---

① G·约翰森已经证实了,随便一个观察者,在区分动画片中以正常重力加速度下落的“落体”动点与“人工”运动时,都不会出错。我不知道是否在婴儿身上做过这样的试验,看一看他们对这些展示是否也是有选择地作出响应。

从事 AI 的人，利用了理论认识方面的这种进展（如果情况是如此的话），能够坦然地忽略学习问题（看起来是如此），直接给执行者**装备**了解决问题时应该“知道”的所有内容。如果上帝一开始就把亚当造成一个有可能解决午夜快餐问题的成人，那么 AI 执行者的制造们**原则上**毕竟也能制造一个“成人”执行者，用世间的知识把它武装起来，**好像**它已经花力气学会了所有需要知道的事情一样。当然，这可能是一种危险的捷径。

这样，装备问题就是一个以这种或那种方式给执行者装备在变化着的世界中做计划所需的全部信息的问题。这是个困难的问题，因为信息必须以能使用的方式来装备。这个问题可以首先分成语义问题和句法问题。语义问题——A·纽厄尔称之为“知识层次”的问题（Newell 1982），就是必须装备何种信息（有关何种主题，得出何种结果）的问题。句法问题就是装入该信息时要使用何种系统、形式、结构或机制的问题。①

在午夜快餐问题的例子中，这种划分是显而易见的。我**列出**了为了解决快餐问题所需要知道的大量平凡事实中的一部分，但这并不意味着我认为这些事实逐一存储于我或任何

---

① 麦卡锡和海斯（McCarthy and Hayes 1969）指出了“认识论的”和“启发式的”这两者之间的不同区分。这种不同表现在，他们把“应当用何种类型的内部标识来表达系统的知识？”这一问题列为认识论问题的一部分（见 P466），而把**那个**句法（因而多少表现为技术细节的）问题从设计“以信息为基础解决这一问题和决定做什么的机制”的过程问题中分离出去。

框架问题究竟是哪种问题，引起这一争论的一个主要原因就是试图对该议题作上述划分。因为对认识论问题所作的句法方面的回答与启发式问题的性质有着重大差别。说到底，如果系统的知识表达所用的句法极其不合常规，那么尽管那种知识表述具备**准确性**，启发式问题仍将是不可能的。同时有些人提出，如果现实中的知识原来就表达得十分得当，那么，启发式问题实际上就不复存在了。

一个执行者之中，其形式是把执行者用得着的那些事实用一长串句子一一清楚地陈述出来。正规地说，这当然是一种可能性，但是将每一个可区分的"命题"分别记入系统的做法，是心理表述的"思维语言"理论的反常的极端形式。没有人赞成这样的观点。即使一本百科全书，也是经过组织，以它的明晰的表达方式而获得重要实用效果的；而活的百科全书——不是为想象中的 AI 执行者画的拙劣的漫画——必须采用不同的系统原理，以获得有效的表述和通路。我们知道成千上万的事情：我们知道刀子在接触蛋黄酱时不会溶化掉，知道面包片比珠穆朗玛峰小，知道打开冰箱不会在厨房里造成核杀伤。

在我们身上，以及任何智能执行者身上，肯定存在着某些高效的、局部生成式的或产生式的系统，用以表述所有需要的信息（存储待用）。这样，我们就必须以某种方式立即存储许多"事实"，并假定这些事实大体上与意义不同的陈述句一一对应地排列着。此外，我们不能指望在实际中得到所谓斯宾诺莎式解答的东西——少量的公理和定义，从这些公理和定义出发，就能推演出所要求的所有其他知识，因为在这些大量的事实之间显然完全不存在任何指定关系。（当我们如我们必须做的那样依赖经验去分辨世界是什么的时候，经验告诉我们的东西，根本不是来自那时我们所知道的内容。）

对于有效信息存储系统的要求，部分地是一种空间限制，因为我们的大脑没有那么大，然而它更重要的是一种时间限制，因为对现实世界中的执行者来说所存储的信息如果不能在通常是有效的短暂实时段中可靠地取用，就根本没什么用处。一个能够在足够长的时间里比如说一百万年间解决任何问题的生物，事实上根本不是智能的。我们生活在有时间压力的世

界之中,就必须能在行动之前迅速地想。(我们不必把这一点看作智能的先决条件。不难看到,我们的思维的确很敏捷,所以就存在着一个我们是如何设法做到这一点的经验问题。)

AI 研究者面临的任务看来是设计这样一个系统,它能使用从它存储的知识中恰当选择的因素来做计划,而这些知识正是关于它在其中运作的那个世界的知识。对于**我们**怎样做计划所作的"内省",提供了如下的过程说明:设想某一情境(往往是概述);然后想象在这一情境中完成某一行动;然后"弄清"所设想的这一情境中行动的可能后果会是什么,并且对它作出评估。可以说,容许这样"弄清"(并使它成为和它本身一样可靠)的发生在幕后的东西,决不可能由内省获得。

在一些较为罕见的场合,我们都经历过这样一番思想体验,思想以审慎的沉思速度在意识中展开。在这些场合中,我们面对某个新的、有一定难度的问题,诸如:怎样把钢琴搬上楼去?用什么方法在不破坏泥制天花板的情况下安装吊灯?我们看到一个人为了解决午夜快餐问题,不得不用**那种**(有意识的和缓慢的)方式进行思考,也许会相当奇怪。但是有迹象表明,即使做计划和身体导向这些我们未曾注意到(虽然从某种意义上说,我们"面对"着它们)的平常问题,也是通过类似过程解决的。这是为什么?因为在这种情境中,我们没有**观察**自己做计划的过程。这一事实足以让传统的、内省的现象主义者相信,这种做计划的过程根本没有进行[1]。另一方面,

---

[1] 这种观察也说服了 G·赖尔,这位举足轻重的内省现象学家(而不是"行为主义者"),见 Ryle 1949。

我们不难想象赖尔对 AI 的反诘:"在制作三明治的过程中,我要做出多少个推理?什么三段论能使我相信啤酒还在杯子里?"有关赖尔怀疑论论点及其与认知科学的关系的进一步论述,见我的"心理表述的方式"(Dennett 1983)。

异质现象学主义者作出这种推理:有关这一情境中对象以及有关候选行动的拟议中的作用和副作用的信息,**必须以某种方式**来使用(来考虑,注意,应用,体察)。这是为什么? 因为否则的话,“巧妙”的行为就完全成了碰运气或变魔术。(有什么模型来说明这种无意识的信息体察是怎样完成的吗? 迄今为止,我们所有的唯一模型就是**意识的**、审慎的信息体察。AI认为,也许这是一个不错的模型。如若不然,我们现在就全都处于黑暗之中了。)

通过考虑反面的事实陈述,我们更加确信一个执行者具有智能:假如有人告诉我火鸡有毒,或啤酒会爆炸,或盘子很脏,或是刀子不结实无法摊开蛋黄酱,我们还会像以前那样行动吗? 假如我是一台愚蠢的“自动机”——或是像泥蜂科黄蜂那样,“不动脑子”地用一成不变的方式重复它的巢穴检查的固定程式,直到累死[①] ——我可以以不恰当的方式“装模作样”地制做午夜快餐,而不理会难对付的环境特点。[②] 但事实上,我

---

① “到了产卵的时候,泥蜂科黄蜂就会专门为产卵建一个窝,同时找一只蟋蟀,把它叮得麻木,但不弄死它。它把蟋蟀拖进它的窝中,紧靠着蟋蟀产下卵,把窝封起来,然后飞走,再也不回来。在适当的时候,卵孵化后,黄蜂幼虫以那只麻木的蟋蟀为食,它没有腐烂,以黄蜂的方式保存,相当于冷冻起来。对人类心灵而言,这样一套精心组织、同时看来颇具目的性的固定程式展现了令人信服的逻辑和周密思考的特征——直到更多的细节被检查为止。例如,黄蜂的这套固定动作是把一只麻木的蟋蟀带回窝边,把它放在窝口,进到里面,得知一切准备停当,钻出来,然后再把蟋蟀拖进去。如果黄蜂在窝内作初步查看时,有人把蟋蟀移开几寸,黄蜂出窝后就会把蟋蟀带回窝口,而不是带进窝里,然后重复进行进窝鉴定一切都准备得当的预备性过程。如果黄蜂进窝时,蟋蟀又被移动几寸,黄蜂会再一次把蟋蟀移回窝口,再进入窝中作最后的检查。黄蜂从未想到直接把蟋蟀拖进窝中。有一回,这一过程重复了40次,每次都是同样的结果”(Wooldridge 1963)。

我在《神机妙算》中讨论了这个昆虫中常见现象的生动例子,(Dennett 1978),也见 Hofstadter 1982。

② 见 Dennett 1982C:58—9,《机器人剧院》。

的午夜快餐制作行为在以多种多样的方式感受有关这一情境的当前的和背景的信息。它能如此感受——进行内隐式的异质现象学推理——的唯一途径是检查或检验问题中的信息。信息处理可能是无意识的、迅速的,它无须(更好是不)包含成千上万的逐一检验过程,但是它必然以某种方式出现,而且当我使自己行动时,它的这些优点必然会及时给我以帮助。

当然,经过几年时间,我可能已经形成了制作午夜快餐的程式,这时我能部分地依靠它来引导我的行动。像这样复杂的"习惯"必然受控于某种复杂性机制,因为,甚至固定步骤的序列也要有定期的检验,以确保次级目标得到满足。即使我是一个偶尔吃快餐的人,我肯定也知道摊开蛋黄酱,做三明治,从冰箱取出东西,这样一些程式,从而可以组成我的多少有点新意的活动。如果把这些程式完美地整合为一体,它们至少能在我更为"不动脑子"的尝试中满足我解决框架问题的需要吗? 这是一个悬而未决的问题,我以后再来讨论。

不管怎样,重要的是从一开始就承认,并经常提醒自己,甚至资质很高的人,也会犯错误。我们不仅不是一贯正确的计划者,而且我们很容易忽视计划中那些重大的、回过来看又是非常明显的缺陷。这种弱点会在常见的"习惯力量"错误的情形中暴露出来(这时,我们的老一套程式表现出对某些奇特环境变化惊人地不敏感,而对另一些则惊人地敏感)。在我们有意识地审慎思考的情况下,也偶尔出现同样的弱点。为搬运钢琴之类的事做计划——在这计划中,你事先已经想过甚至"走过"整个运作过程——不过是为了发现这种情况:某个完全可以事先看到,但却未看到的障碍物,或者未曾想到的附带情况冒出来时,你必须退回去或放弃原来的计划,这种事你

遇上过多少次？如果我们这些精明人实际上难得使自己陷入困境，这并不是因为我们事先计划得那么好，而是因为我们在工作过程中，把回忆起来的经验知识（例如关于傻子使自己陷入困境的经验）和进展中的经常性检查结合起来，以弥补我们蹩脚的计划能力。即使如此，我们必须有足够的知识才能在恰当的时候回忆起恰当的经验知识，以及确认那些即将出现的问题。

我们作一总结。经过条理清晰、理由充分的思考之后，我们的结论是：智能执行者必须着手制订感受信息敏捷的"计划"，其作用是对他的行动后果产生可靠的而不是过于简单的预期。在智能动物中，这些预期一般是有效的，这一点可以从预期受挫时它们表现出的吃惊反应得到证明。这为我们刻划能生成框架问题的最低目标的特征提示了一个生动的方法：我们要让制作午夜快餐的机器人对玩魔术用的盘子、摊不开的凝固的蛋黄酱以及我们将啤酒杯粘在架子上的事实感到"惊讶"。为了感到惊讶，你必须预期别的什么事物，而要预期某个别的正确事物，你必须具有**并使用**大量有关现实中事物的信息。①

---

① H·德雷福斯曾经指出，**不预期** x 并不意味着**预期** y（这里 $x \neq y$），所以一个人因某个并未预期的事物而受惊时，他并不一定（无意识地）预期了某个别的事物。但是这种**不预期**的意识并不足以解释为什么会受惊。在此后 5 小时里你迟早会看到阿尔法·罗密欧、别克、雪佛莱和道奇这四辆汽车按字母次序停放，这种事的难度怎么样？无疑，所有的事都考虑到，这太难了，所以我不会预期你去预期此事；同时我也不会预期你会因看到这种未预期的景象而受惊——除非在某种特定场合，你有理由在这一时间和地点预期某个别的事物。

受惊反应是认知状态的有力的指示器——警察们（还有侦探小说作者们）早就知道这一事实。只有某个预期冰箱里装着（比如说）史密斯尸体的人，才会在看到里面装着不大可能的三件东西：一瓶夏布利名酒、一筒猫食和一块洗碟布时**受惊**（而不是多少有点兴趣）。

有些人从预期的主要作用中得出这样的结论:框架问题根本不是一个新问题,它与做计划的行动毫无特殊关系。在他们看来,这只是一个对任何未来事件做正确预期的问题,不管这些事件是一个人自己的行动,还是另一个执行者的行动,或者只是自然界中偶然发生的事情。这是一个归纳的问题,休谟提出它,古德曼(Goodman 1965)对它作了进一步阐发,但是仍没有人人都满意的解答。现在我们知道,归纳问题的确是个棘手的问题。主观概率和信仰定位的理论并未在思考的平衡状态中保持稳定,所以公允地说,还没有人曾经对这个一般性问题作出正确的、原则性的回答,因为如果我相信这一情况的所有方面都是真的(具有这一情况证据的所有方面),那么,(关于未来,或关于现实世界中尚未检验的那些部分)我应该相信的还有什么呢?

把一个尚未解决的问题简化为另一个问题,尽管还有不能令人满意的地方,却是某种进步,然而它不是这种情况下的选择。框架问题不是乔装打扮后的归纳问题。因为我们假定归纳问题已得到解决。假定——这也许是不可思议的——我们的执行者已经解决了它的所有归纳问题,或者已通过命令使它们得到解决;于是它相信来自证据的所有正确概括,并把恰当概率和条件概率同所有这些概括联系起来。按照假设,这个执行者相信的只是关于它知识范围中的所有经验事物,包括未来事件的概率,这些它所应当相信的事物。这对框架问题可能仍不是件好事情,因为这个问题关系到怎样表述(所以才能使用)所有来之不易的经验信息——该问题的出现并不依赖于它具有的任何真值、概率、被认可的可断言性,或主观确定性。即使你具备了关于这个变化着的世界的精良知识

（而不只是信念），这一知识怎样才能表述得使它能被有效地应用呢？

回想一下可怜的 R_1D_1，并且为证论方便起见，假定它在测出它的所有行动的全部后果的概率方面有着完善的经验知识，所以它相信，执行拉出（小车，房间）会使小车车轮发出听得见的噪声的概率是 0.7864；通向房间的门向里开而不是向外开的概率是 0.5；房间里没有活着的大象的概率是 0.999996；车子移动时炸弹还留在车上的概率是 0.997。R_1D_1 怎样才能最终从经验知识的大海里捞出那根意义攸关的针呢？一本活的百科全书由于它有关于悬崖和重力作用的全部知识会轻蔑地对待悬崖，除非它被设计得能在恰当的时候找到适当的知识，因而能够做出它与现实世界交往的计划。

最早的 AI 计划系统研究采用的是演绎方法。鲁宾逊创造的分解定理证明法，对设计者们是一个鼓舞，他们希望把系统具有的所有"关于世界的知识"明晰地表述为公理，并使用普通逻辑即谓词演算来推演行动的结果。设定某一情境 S，就是使系统包含一组描述这一情境的公理而为其建立模型。再在其上增加背景公理（即所谓"框架公理"，框架问题因此而得名），它们描述了为该系统定义的每一行动类型的一般条件和一般结果。该系统将行动施加于这组公理——假定在情境 S 中有某个行动 A 出现——然后推演出 A 在 S 中的结果，于是就得到关于结局的情境 S′的描述。虽然在我们的意识经验中，所有这些逻辑演绎好像根本不存在，可是演绎方法的研究得以进行是根据下面两个有力假定之一或两者兼有之：心理实在论是"纯"AI 的无根据的额外添加，而不是它的目标——方法论假定；所描述的演绎过程能以某种方式为意识途

径之外的幕后过程建立模型——存在性(即使仍是模糊的)假定。换句话说,或者我们在谓词演算中不用演绎方式进行思维,但机器人会用;或者我们在谓词演算中确实(无意识地)用演绎方式进行思维。然而,撇开对心理实在论的怀疑不说,除了一些人为造成的琐碎例子外,演绎方法还未曾表现得有效,有效——这是对每一机器人优劣的实践证明。

我们来看几个与行动类型“**把 x 移到 y 上**”有联系的典型的框架公理:

(1) 如果 $z \neq x$,同时我把 x 移到 y 上,那么如果 z 事先在 w 上的话,事后 z 仍在 w 上。

(2) 如果 x 事先是蓝色的,同时我把 x 移到 y 上,那么事后 x 仍是蓝色的。

我们注意到,关于保持蓝颜色的(2),只不过是我们必须与这一行动类型相联系的众多烦人的“无变化”公理的一个例子。更糟的是,我们注意到:也是关于保持蓝色的(2)的同类们,必须与其他每一个行动类型相联系,例如与**捡起 x** 以及与**把 x 给 y** 相联系。人们无法一劳永逸地通过作出类似。

(3) 任何蓝色物体仍保持为蓝色的假定来避免这个无知的重复,因为这样做是错误的,特别是我们要为引入像“**把 x 漆成红色**”这样的行动类型留余地。既然事实上一个情境的任何方面都可能在某种环境中发生变化,这个方法就需要为每一方面(关于 S 的描述中的每一论断)引入一个公理,以处理每一行动类型是否在这一方面发生变化的问题。

这种表述上的无度发展很快就变得无法控制,但是对于 AI 中某些“玩具型”问题来说,混合使用玩具型环境和蛮力方式,可以在一定程度上将框架问题制服。SHAKEY 的早期形

式,SRI 的机器人,就是在这种简化而单调的世界中运作的,它要操心的方面很少,所以有可能侥幸地完成对框架公理的彻底考察。[①]

防止这种公理数目激增的尝试以这一建议作为开始:系统的运作根据的是一个不言自明的假定,即除了所施加行动的定义中明确断言要发生变化的东西以外,一个情境中的任何东西都不发生变化(Fikes and Nilsson 1971)。正如 G·哈丁曾注意到的那样,这里的问题是,有一件事你正好没有做。这是 R_1 在没有注意它会把炸弹同小车一起拉出来时所遇到的问题。在对我的午夜快餐解答的明确表述中(几页前),我提到了把盘子拿来放在桌上。根据这一建议,我的 S′模型将把火鸡留在厨房里,因为我没有明确地说火鸡会随盘子一起来。人们当然可以通过修补“取”或“盘子”的定义来处理这一问题,但这只有以增添别的定义为代价。(多一些补丁就能解决这一问题吗?人们应当在什么时候放弃打补丁而去寻找一个全新的方法?这正是在该领域中经常碰到的方法论上的不确定性,同时当然也没有人能负责地事先声称有处理这些事务的良好规则。失望中的幼稚建议和对变革的呼唤,与固执地坚持走毫无希望的道路显然是同样应当避免的。难怪这是一个有争议的领域。)

虽然我们不能侥幸地采取这种策略:假定“一个人能做的只是一件事”,但是在任何情境中(逻辑上)可能发生的事情的

---

① P·海斯指出的 SHAKEY 的这种早期特点,引起过我的注意。也见德雷福斯的文章(Dreyfus 1972:26)。在丹尼特的文章中 SHAKEY 的作用全然不同(Dennett 1982b)。

确很少发生，也同样是事实。有没有什么方法给可能出现重要副作用的范围加上容易出错的标记，并假定该情境的其余部分保持不变？在这里，进行相关性检验看来是个好主意，这些检验可能确实有这个作用，但它们不属于演绎方法。正如明斯基注意到的那样：

> 即使我们对相关性限制作出系统阐述，逻辑系统使用它们时也有问题。在任何一种逻辑系统中，所有公理必然都是“容许的”——它们全都有助于容许新推论的推出。每一个新增的公理都意味着更多的定理，没有一个定理会消失。绝对没有一个直接增加信息的方法，是就**不**应该推出的种种结论告诉我们一个如此系统的！……如果我们试图通过增加相关性公理来改变这一情况，我们还会得到所有那些并不想要的定理，还有那些令人厌烦的关于它们的不相关性的陈述(Minsky 1981:125)。

我们所需要的是一个真正**忽略**它所知的大部分内容，并在任一时刻根据恰当选择的部分知识进行运作的系统。恰当选择，并非是经过彻底考察作出的选择。可是，怎样才能为一个系统建立忽略**规则**？或更进一步，既然明确地遵循规则不等于这个问题，怎样才能设计出一个系统，使它在一个复杂的行动环境中面临种类众多的不同情况时，可靠地忽略那些应该忽略的东西？

J·麦卡锡称这为限定性问题，并且通过著名的教士和食人者难题对此作出生动说明。

三个教士和三个食人者来到河边。只有一条可乘两人的小船。如果在河岸的任一边食人者的数目超过教士，教士就会被吃掉。他们该怎样过河呢？

显然，该难题需要的是设计出一个往返划船策略，使他们都能过河，又不发生惨剧……

我们设想把这个问题交给某人，经过一段时间的苦思冥想，他提出建议，向上游走半英里路，从桥上过去。"什么桥？"你说，"这个问题的陈述中没有提到过桥。"这个笨伯回答说："可是他们没有说那里没有桥。"你查看了一下原文，甚至查看了翻译成一阶逻辑的原文，你不得不承认，"他们没有说"那里没有桥。于是你修正了问题，将桥排除在外，再次提出。这个笨伯又建议乘直升飞机，当你排除直升飞机之后，他建议飞马，或是在两人划船时，其他人吊在船外。

现在你看到了，他虽然是个笨伯，但却是一个有创造性的笨伯。要让他正确认识这一个难题的实质，你完全失望了，于是把解答告诉他。更让你气恼的是，他以船可能会漏或没有桨为理由，对你的解答提出反诘。当你纠正了该问题陈述中的这一省略之后，他又提出可能有一个水怪游出河面，把小船吞下去。你再次受挫，于是去寻找一个能一劳永逸地使他哑口无言的推理方式(McCarthy 1980:29－30)。

正常的智能人在这种情境中所做的事，是进行某种形式的非单调推理。在经典的、单调的逻辑系统中增加前提时，决不会因这些前提而减少所能证明的内容。正如明斯基所指出

的，公理基本上是容许性的，而一旦一个定理被容许，增加更多的公理决不会使先前定理的证据无效。但是当我们考虑一个难题或实际生活中的问题时，我们可能会得到一个解答(甚至证明它是一个解答，而且是该问题的唯一解答)，然后发现由于在提出该问题时增加了一个新因素，我们的解答变得无效了，例如："我忘了告诉你，那里没有桨"，或是"顺便说一句，在上游有一个很结实的桥"。

这种事后增补使我们看到，与我们的假定相反，其他事项并不是同等重要的。我们一直是在借助**其余情况相同**的假定进行推理，现在发现了某个"反常"事物的存在，我们的推理则因此而遭到损害。(顺便提及，这里所说的反常性，比起任何人根据概率论苦苦得出的任何东西要微妙得多。正如麦卡锡指出的，"涉及带有假定特性的食人者的整个情境，不能被看作是具有概率的，所以一旦给出该假定，很难认真对待桥的条件概率。")(出处同上)

存在于一些推理之中的**其余情况相同**条款有其优点，即人们不必确切地说出它指的是什么。"你说'其他事项等同'，这是什么意思？确切地说，何种其他事项的何种格局被看作等同的？"如果必须回答这样的问题，援引**其余情况相同**条款是毫无意义的，因为人们正是为了回避这一任务才使用它的。如果能够回答这个问题，就不需要在一开始援引这一条款了。因而考察框架问题的方法之一，正是尝试让计算机利用这种具有人类特点的心理操作方式。AI 中当前正在研究的非单调推理方法有好几种，它们的差别很大，只有这个目标是共同的：获得人类的才智，去忽略该忽略的，同时在重大的反常情况出现时保持警觉。

以 M·明斯基和 R·尚克(Minsky 1981；Schank 和 Abelson 1977)的工作为典型的一组方法,从集中注意能力的定型套路中获得了忽略的能力。在这里,这个富有启发性的见解表现为这种思想:将一切变化多端的生活经验压缩成数目可控制的一些定型套路的主题,亦即范式型的情节梗概——明斯基称之为“框架”,尚克称之为“脚本”——中的变化。

一个人造执行者配有内容充足的框架或脚本的纲要,如果这些框架或脚本恰当地相互联系,并通过执行者的感知器官与现实世界的撞击恰当联系,这个执行者就会运用一个详细描述的系统去面对现实世界,组成这个系统的东西可以称之为习惯注意和有利倾向,即在特殊类型的境况中迅速得出特殊类型结论的倾向。它会在某些环境中“自动”对某些特征加以注意,并假定某些通常有的、这些环境中未检查出的特征是存在的。与此同时,它有区别地对重大偏差保持警觉,这些偏差背离了那些它从开始就一直“预期”着的定型套路。

对这种执行者与现实世界碰面的片断所作的模拟表明,在许多情境中执行者都做到了行止得当,而且看上去很自然,当然很难说这种方法的局限性是什么。然而怀疑主义也有充分的根据。最明显不过的是,虽然当现实世界与系统的定型套路相配时,甚至与预期的定型套路的变化相配时,这种系统的表现是靠得住的,但是当现实世界变得反常时,这种系统一般都不能得体地从它们被引入的错误分析中复原。事实上,**在极端情况下**,系统行为寻求的只是某一种现实世界,就像昆虫在它们刻板的向性和其他由遗传硬连接的行为程式的错误引导下做出的一些愚蠢地产生相反结果的活动。

当这些令人为难的意外情况发生时,系统设计者可以用

增加条款的方法改进设计，以应付这些特殊情况。重要的是应看到，在这些情况下，系统并不重新设计自己(或学习)，而是必须等待外部设计者选择改进的设计方案。这一重新设计的过程从某种意义上说是对自然选择过程的简要描述，它倾向于作出最低限度的、零星的专门的重新设计，有点像在未来事件可能存在的模式上下赌注。所以从某种意义上说，它是符合生物学主题的。[①] 然而，这种系统在具备从自己错误中学习、而不用设计者干预的重要能力之前，它将继续以昆虫的方式作出响应，以这种行为方式作为人类日常生活中的反应模型，是远不切合实际的。依赖定型套路所提供的捷径和取巧方法，在人类思维方式中是够明显的，但是同样明显的是，我们具有更深层的理解，可以在捷径无效时退回，所以，将某种程度的这一更深层的理解植入系统中，看来是使该系统快速和得体地学习的必要条件。

实际上，脚本或框架方法是**预先解决**特殊执行者有可能遇到的框架问题的一种尝试。虽然昆虫看来的确装备了这种控制系统，可是人却不同，即使他们的确表现出对定型套路的依赖，却仍有做替补的思维系统，能以更强有力的方式处理所出现的问题。再者，人在利用定型套路时，依靠的至少是自己设计的定型套路，并且时至今日，还没有人能够提供什么切实可行的思想来说明，以前的经验会怎样指导一个人的框架生成或脚本编写机制。

近年来，从演绎传统中产生出若干个不同的精心策划的

---

① 然而从一个重要的方面来看，它与自然选择过程有着显著的不同，因为该过程的试错和选择决非是盲目的。但是也会出现这种情况：在重新设计的过程中，缺乏耐心的研究者只不过是靠着先见之明的干预，来缩短时间而已。

尝试，为更深层的理解提供了表述框架。D·麦克德莫特和J·多伊尔创造了“非单调逻辑”(McDermott and Doyle 1980)，R·赖特提出“缺省推理逻辑”(Reiter 1980)，而J·麦卡锡发展形成了“划界”系统，这是一个形式化的“猜想规则，人或程序可使用它来‘直达结论’”(McCarthy 1980)。它们之中没有一个是，或者自称是，**其余情况相同**推理问题的全解，但是它们或许可以作为这种解的组成部分。最近，麦克德莫特提出了“供有关过程和计划的推理用的时态逻辑”(McDermott 1982)。我不打算分析这些方法在形式上的长处和缺点。我主要有另外一种担心。从一个角度来看，非单调逻辑或缺省逻辑、划界、以及时态逻辑，对于不动脑子的机械的演绎方法来说，看来都是根本性的改进；但是换一个稍微不同的视角，它们则表现出更多的共同之处，至少与用于心理模型的框架一样地非实在。

它们以先前的姿态出现，是要向更高的心理实在论前进一步，因为它们认真对待并试图表述的以现象学方式凸现的现象，是属于常识的**其余情况相同**“直达结论”推理的。但是意识思维的幕后实现方式**在人类中**是怎样完成的，它们真的成功地就这一点提出了什么有说服力的见解吗？即使在大有希望的未来某一天，一个带有可排除错误的划界法的机器人，在非玩具型环境中操纵自如，它的那些**在低于现象学的层次上**描述的组成过程，与人类之中那些未知的较低层次的幕后过程具有信息上的联系，这种可能性是否很大呢？为了更清楚地说明我的担心，我打算引进**认知之轮**的概念。

我们先想想普通轮子，就能够理解认知之轮可能是什么。轮子是奇妙而精巧的技术成就。对于传说中的轮子发明者的一贯尊崇，完全是理所应当的。可是既然轮子这样奇妙，为什

么没有带轮子的动物呢？为什么在自然界中没有发现轮子(或像轮子一样的功能)呢？首先，必须确认这一问题中的假定是否正确。几年前，从几个微小生物(一些细菌和一些单细胞真核生物)身上得出一个令人震惊的发现：它们具有轮子一类的东西。它们那起着推进作用的尾部，一直被看成柔性的节鞭，其实是有一定硬度的螺旋体，在具备主轴承的微型发动机之类的东西推动下连续转动。[①] 人们更了解的是风滚草，不过由于明显的原因，较少引起人们的兴趣。所以自然界里没有轮子(或轮式设计)的说法是不符合事实的。

可是，宏观的轮子——爬行动物、哺乳动物或鸟类中的轮子——尚未发现。为什么没有呢？例如，对有些鸟类来说，轮子似乎是绝好的可收缩的降落装置。这个问题一经提出，就会涌现出一些看似合理的、可解释它们不存在原因的理由。最重要的可能是有关轴和轴承交界处的外形特征的考虑，因为这些特征使物质或能量穿越交界处的传递特别困难，生命系统中维持生存的大动脉怎样才能保持完整地穿过这个交界处呢？但是一旦提出问题，解答就会露出端倪。假若有生命的轮子以不旋转、无功能的形式长到成熟，然后变硬脱落，但不完全脱落，就像鹿角或快速长成的虫壳那样，然后它就在滑润的固定轴上自由旋转。这可能吗？很难说。它有用吗？也很难说，尤其因为这样的轮子必须靠惯性滑行。这是一个有趣的推测练习，但肯定不会促使我们得出明白无误的先验结论。断言轮子在生物学上是不可能的，显得有点鲁莽，但我们同时也能体察到：对自然界的设计问题而言，这解答至少还是

---

① 有关这些论点的详情和进一步的见解，见 Diamond 1983。

很遥远的，而且未必能成功。

目前，认知之轮只不过是认知理论中某种远非生物学的设计建议（可处在从最纯粹的语义层次到最具体的神经元“线路图”层次的任一层次上），不管它作为一种技术是多么奇特和精巧。

这显然是一个定义模糊的概念，只有作修辞简化的用处，像是一种指示出需要仔细加以研究的实际困难的手势。“要当心对认知之轮的假定”，这话看起来像是对认知研究者的忠告，实际上是把虚情假意的空洞内容当成了要遵循的准则。[①] 它与证券经纪人“低买高卖”的准则有异曲同工之处。其实，该术语用来确定讨论主题是相当不错的。

许多批评 AI 的人都深信：**任何** AI 系统除了作为认知之轮的齿轮箱之外，什么作用也不起，而且这是必然的。当然，这有可能是正确的，但是所以有这种看法，一般来说是基于对这一领域的方法论假定的误解。有关某个认知现象的一个 AI 模型被提出时，这个模型可以在多个不同层次上作出描述。从整体性最强的现象学层次开始，在这一层次上，行为是通过普通的心灵主义术语（不无牵强地）来描述的，向下通过各种实现方式的层次，直到程序编码层次——甚至进一步向下，到达基本的硬件操作层次，这取决于人们的意愿。人们不会假定这个模型**一直向下**映射到心理学和生物学过程。该主

---

① 我饶有兴趣地发现，至少有一位 AI 研究者在刚一听到我这个新术语时，就对它的修辞含义发生了误解。他把“认知之轮”看作一种赞词。如果人们像他一样，不是把 AI 当作一种心理学研究方法，而是看作企图扩展人类认知力量在工程方面的一个分支，那么认知之轮当然就成了惊人的发现。计算机有庞大的而且事实上永不会出错的记忆能力，也许是最重要的例子；还有，计算机具有精湛的算术技艺和无懈可击的不知疲倦、不会注意力分散的特点。见霍施塔特（Hofstadter 1982）就疲倦与记忆结构和创造力条件的关系所作的富有洞察力的论述。

张仅仅是说，在某个高层次上或是在现象学层次(该层次仅仅提出问题)以下的一些描述层次上，存在着从模型特征到被建模型事物的映射，这也是人或其他生命体中的认知过程。可以这样理解，在拟建立模型的层次之下，所有的实现方式细节无疑都可以由认知之轮构成，这是一些非生物的计算机的活动，它们从整体效果上对认知的次级组成部分进行模拟，所使用的方法完全不同于有待于在大脑中发现的方法。有些人没有意识到，由认知之轮在微观上构成的模型，在较高的聚合层次上，仍能与生物过程或心理过程达到有成效的同构，他们认为关于AI的普遍怀疑主义有其充分的先验理由。

然而，承认可能存在着有价值的建立模型的中间层次，并不能确保它们的存在。在一个特殊例子中，一个模型有可能从现象学上可识别的心理描述层次直接下降到认知之轮的实现方式，而根本不解释我们人类是怎样设法利用这种现象学的。我怀疑，在这一领域中，当前所有涉及框架问题的建议都有这种不足。也许应当排除先前那个只不过是自述式的命题。我感到很难想象(它具有什么价值)，任何机械形式的过程细节，比如属于麦卡锡划界法的，会在尚待讲述的有关人类的常识推理是怎样完成的幕后过程中有合适的对应物。如果这些过程细节缺乏“心理实体”，那么在该建议中就没有留下什么可能为心理学过程建立模型的东西，除了利用直达结论、忽略以及类似的方法所做的现象学层次的描述——我们已经知道我们正是这样做的——之外。

然而，对于这样一些理论上的探索有另一种辩护，我认为应当认真地对待它。有人主张(这里采用麦卡锡的主张)，虽然麦卡锡方式的形式化常识推理没有直接告诉我们有关心理

推理过程的什么事情，但是它澄清了对所要求的任何一种像这样的生物或非生物实现方式的纯语义层次的特征，并使之明确化、系统化。当我们认真地把信息加工看作时空之中的真实过程而向前迈出一大步之后，可以再向后退一小步，探讨一下这个十分抽象的层次的进步意义所在。即使在这一颇为形式的层次上，划界法以及其他非单调推理文本的功能仍然是个有待解决的、但显然是可以探索的问题。[①]

有些人做过这样的考虑：在涉及句法层次的实现方式之前，必须要有对语义层次的背景作完整的再思考，这是较为现实地解答框架问题（而且实际上十之八九是任何解答）的关键。在表述"可信命题"时为了填充谓词演算格式，而挑选出来的一批基本上标准的谓词和关系，有可能实现的是对这一任务来说在性质上多少有些不恰当的语法分析。在大多数情况下，对这些系统中的公式的解释，使得世界不再沿着那些熟悉的路线即物体在时间和地点上具有某些特性的路线行进。关于世界中的情境和事件的知识是由也许可称为词语快射（verbal snapshot）系列的东西来表述的。状态 S 由 t 时刻的真语句表从构造上加以描述，而这些语句断言各种详细情况的各种 n 目谓词为真。S 转换成状态 S′，在 t′时刻也有类似的真语句表。根据历程和过程重新构思"做计划的世界"或许会更好一些吗？[②] 有一些原理适用于由术语（在本质上是事物

① 麦克德莫特文章（"过程和计划推理的时态逻辑"，§6，"实现方式概要"，McDermott 1982）给人以深刻印象，它表明一旦人们着手处理实现方式的问题，会有多少新的论点出现，以及纯形式思考是多么间接（尽管还是有用的）。

② P·海斯一直在探索这一主题，初步论述可见他的"朴素物理学 1：液体的本体论"（Hayes 1978）。

的名称)和谓词表达的、穿越若干时间截面的知识,我们并不尝试根据这些原理为记录事物的能力建立模型,因为我们也许能更直接地建立记录事物的模型,并使所有逐时地被信以为真的东西的截面信息仅仅隐含于格式中(很难将其抽取出来——正如我们遇到的那样)。这是一些诱人的设想,但就我所知,到目前为止,它们还处在变化不定的范围中。[①]

另一个可能与之有关的尚在变化的主题是,当前随框架问题而出现的困难起源于由冯·诺伊曼式串行处理构造体系所产生的概念式构架,这是 AI 中至今仍在使用的计算机构造体系。大型、高速并行处理器的发展影响带来了概念的重大革新,当然这种革新目前还仅仅处在朦胧的设想阶段。大脑毫无疑问是巨大的并行处理器,设想由这种新硬件产生的概念会很容易地改写为现实的心理模型,这是很诱人的。但是谁能证明这一点呢?眼下,乐观主义者关于并行处理功能的大多数主张,与神经科学家工作中常常遇到的简易观测属于同一类东西,神经科学家们假定神经系统的各部分有惊人的认知功能,而没有提供这些功能是怎样实现的线索。[②]

在作为现象的魔术表演和已充分认识的小块大脑组织的功能之间有一片空白地带,将其详尽地填补起来,是一项十分

---

① 我想 O·塞尔弗里奇(O·Selfridge)即将出版的专著《跟踪追迹》(Bradford Books/MIT Press),有希望推进这一前沿领域,但是我尚未能领会他的思想主旨。关于这一主题,即将由 Bradford Books 出版的 R·G·米利肯(Millikan)的《语言、思维和其他生物学范畴》中也有一些富有启发性的章节。

② 以认知之轮为出发点的"由上至下"的理论家有其弱点,与之相对应的是,"由下至上"理论家有发现奇妙组织(wonder tissue)的偏向。(很多地方都有奇妙组织。例如,J·J·吉布森的知觉理论看来就是把整个视觉系统当作一整块奇妙组织,以奇异的感受性,对于大批复杂的"给予过程"作出反应。例如见 Gibson 1979。)

艰巨的研究任务，它摆在未来每一派理论家的面前。但是在能够解决这些问题之前，理论家们必须与它们碰面，而为了与这些问题碰面，他们必须坚定地跨入这片空白地带，提出"怎样做"的问题。哲学家们(以及每个别的人)都很清楚，人——无疑还有所有智能执行者——能够从事快速、灵敏、有风险但有价值的**其余情况相同**推理。人们是怎样做到这一点的呢？AI可能还没有令人满意的解答，但是它至少已经与这个问题碰面了。①

参考书目

Cherniak, C. [1983]. 'Rationality and the Structure of Memory.' *Synthese* [57: 163-86].
Darmstadter, H. (1971). 'Consistency of Belief.' *J. Philosophy* 68: 301-10.

---

① 我发现了几篇看来对思考框架问题有贡献的哲学文章——虽然用的不是框架问题的术语，其中有一篇是R·德苏萨的"情感的理性"(de Sousa 1979)。在题为"情感的作用是什么?"的章节中，德苏萨以令人信服的理由指出：情感的功能是填补由(单纯的需要加上)在确定行动和信念时的"纯粹推理"遗留下来的空白。想一想埃古是怎样蓄意挑起奥赛罗嫉妒的。他的任务主要就是引导奥赛罗的注意，启发他提出问题……一旦注意力按他的指示进行，以前在同样的证据面前甚至未曾想过的推理，就被迫地感受到了。按照德苏萨的理解，"情感是一些具有确定作用的模式，它使事物从注意对象中凸显出来的，是提问的路线，是推理的策略"(第50页)，同时情感不能以任何方式"简化"为"明确表达的命题"。虽然这说法富于启发性，但是它当然没有就怎样使内部(情感)状态具有这些有趣的功能提出什么具体的建议。另一篇富于启发意义而往往被忽略的文章，是H·达姆施泰特的"信仰的一致性"(Darmstadter 1971:301—10)。达姆施泰特探索性地研究了**其余情况相同**条款，以及有可能存在于作为心理状态的信念和信念持有者可能说出的(或是就这些状态已经说出的)句子之间的关系。这些研究提出了许多值得进一步深入探讨的见解。

Dennett, D. C. (1978). *Brainstorms*. Cambridge, Mass.: MIT Press/Bradford Books.
—— (1982*a*). 'Why Do We Think What We Do About Why We Think What We Do?' *Cognition* 12: 219–27.
—— (1982*b*). 'How to Study Consciousness Empirically; Or Nothing Comes to Mind.' *Synthese* 53: 159–80.
—— (1982*c*). 'Beyond Belief.' In A. Woodfield (ed.), *Thought and Object*, pp. 1–96. Oxford: Clarendon Press.
—— (1983). 'Styles of Mental Representation.' *Proc. Aristotelian Soc.* 83: 213–26.
Diamond, J. (1983). 'The Biology of the Wheel.' *Nature* 302: 572–3.
Dreyfus, H. L. (1972). *What Computers Can't Do*. New York: Harper & Row.
Fikes, R., and Nilsson, N. (1971). 'STRIPS: A New Approach to the Application of Theorem Proving to Problem Solving.' *Artificial Intelligence* 2: 189–208.
Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston, Mass.: Houghton-Mifflin.
Goodman, N. (1965). *Fact, Fiction and Forecast*, 2nd edn. Indianapolis: Bobbs-Merrill.
—— (1982). 'Thoughts Without Words.' *Cognition* 12: 211–17.
Hayes, P. J. (1978). 'Naïve Physics I: The Ontology of Liquids.' Working Paper 35, Institute for Semantic and Cognitive Studies, Geneva.
—— (1979). 'The Naïve Physics Manifesto.' In D. Michie (ed.), *Expert Systems in the Micro-Electronic Age*, pp. 242–70. Edinburgh: Edinburgh University Press.
Hofstadter, D. (1982). 'Can Inspiration be Mechanized?' *Scientific American* 247: 18–34.
McCarthy, J. (1968). 'Programs with Common Sense.' *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*, London. Repr. in M. Minsky (ed.)., *Semantic Information Processing*, pp. 403–18. Cambridge, Mass.: MIT Press.
—— (1980). 'Circumscription—A Form of Non-Monotonic Reasoning.' *Artificial Intelligence* 13: 27–39.
—— and Hayes, P. J. (1969). 'Some Philosophical Problems from the Standpoint of Artificial Intelligence.' In B. Meltzer and D. Michie (eds.), *Machine Intelligence 4*, pp. 463–502. Edinburgh: Edinburgh University Press.
McDermott, D. (1982). 'A Temporal Logic for Reasoning about Processes and Plans.' *Cognitive Science* 6: 101–55.
—— and Doyle, J. (1980). 'Non-Monotonic Logic.' *Artifical Intelligence* 13: 41–72.
Millikan, R. G. [1984]. *Language, Thought and Other Biological Categories*. Cambridge, Mass.: MIT Press/Bradford Books.
Minsky, M. (1981). 'A Framework for Representing Knowledge.' Originally published as MIT AI Lab. Memo 3306. Quotation drawn from excerpts repr. in J. Haugeland (ed.), *Mind Design*, pp. 95–128. Cambridge, Mass.: MIT Press/Bradford Books.
Newell, A. (1982). 'The Knowledge Level.' *Artificial Intelligence* 18: 87–127.
Reiter, R. (1980). 'A Logic for Default Reasoning.' *Artificial Intelligence* 13: 81–132.
Ryle, G. (1949). *The Concept of Mind*. London: Hutchinson.
Schank, R. C., and Abelson, R. P. (1977). *Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge*. Hillsdale, NJ: Erlbaum.
Selfridge, O. (forthcoming). *Tracking and Trailing*. Cambridge, Mass.: MIT Press/Bradford Books.
de Sousa, R. (1979). 'The Rationality of Emotions.' *Dialogue* 18: 41–63.
Wooldridge, D. (1963). *The Machinery of the Brain*. New York: McGraw-Hill.



朴素物理学宣言

P·J·海斯*

花园里依然可见那长满杂草和黑甘蓝的角落，还有乱石和花梗。房间有的地方灼热，有的地方阴冷，有许多黑色的孔洞和嘤嘤作响的木牌，还有让人敬畏的祭坛之地，它依傍着那片一望无尽的土地，上面布满小东西和装饰物，它们卷曲着，紧固着，时而喧闹，时而叹息，时而打开，时而又关上。

摘自 L·李："玫瑰色的苹果酒"

1. 导　　言

人工智能充满了"玩具型问题"：或是小型的、人为的公理化，或是设计出种种难题，用来训练各种问题求解程序、表述性语言或系统的才干。这些受训者极需要一些非玩具型的、可以进行实验的实际领域。在这篇文章中，我提出对有关日常物理世界的相当大的一部分常识性知识，即对关于物体、形状、空间、运动、物质(固体和液体)、时间等的知识加以形式

化的构造方案。

下面我将概述这一方案，并使之区别于另一些表面上相似的方案；就出现的某些普遍性问题进行讨论；阐明这方案需要实现和能够实现；然后简述使之实现的方法。沿着这一思路，我将简要说明这方案所假定的意义理论，并对另外一些理论作出批评。

2. 方案总览

这一方案要为物理世界中大量存在的普通日常知识构造一个形式化。举例来说，这种形式化有可能是一阶逻辑形式体系中若干推断的集合，或是 KRL“单元”的集合，或是一个微型规划程序，或是许许多多别的什么东西中的一个。本方案并不打算建立新的形式体系来表达这种知识。尽管我们承认，形式体系也许在适当的时候有必要作大刀阔斧的改进，但是我们相信，已有的、大家熟知的形式体系尚有许多未曾发掘的潜力。本方案不打算用形式体系的方式编写能够解决问题、制定行动计划或做其他事情的程序。我有意地忽略了对执行过程细节的考虑，虽然控制和搜索问题，简单说就是**计算问题**，是必须重视的。我们一再看到，由于过早地考虑计算问题，在 AI 的表述问题方面的严肃工作被转移方向或全面受阻。

---

* P·J·海斯“朴素物理学宣言”一文，选自 D·米基编辑的《微电子时代的专家系统》(爱丁堡大学出版社，1979)，第 242—270 页。

帕特里克·J·海斯(Patrick Hayes)，施乐公司研究中心的高级科学家。

我们认为形式体系应具备以下特征(详细说明见后):

1. **彻底性** 它应当涵盖日常物理现象的整个范围,例如,不仅仅局限于积木世界。既然从某个重要的角度来看,世界(甚至日常世界)包含着无限丰富的潜在现象,所以这种彻底性永远不会完美无缺。然而,我们应当尽量把所有重要的空位填满,至少应确认其存在。

2. **精确性** 应当对它作详细推敲。例如,在对积木世界的描述以及支撑关系中,一块积木的诸方面,如形状、材料、重量、刚度和表面组织,都应是可获得的概念。同时,由于世界可以无限细分,完全精确是不可能的,但是比起普通"玩具型问题"公理化那种十分粗劣的精确性,我们应当尽力做得更好些。例如在这种公理化中,一块积木位于另一积木"上方"的关系只是一个局部编序,这样才可能以整数作为公理的模型。

3. **稠密性** 事实与概念之比必须相当高。换句话说,这些单元必须含有大量接口。从某种意义上说,低密度形式化是没有价值的,因为它们无法对所含的概念作出充分说明,以便完全确切地将意义固定下来。有时,为了特定的目的,例如进行基础研究,低密度可能有其优越之处,但对我们来说却不然。

4. **统一性** 对于整个形式化来说,应当有一个共同的形式框架系(语言、系统等),这样,不同部分(公理、框架……)之间的推论联结才能显而易见;同时子形式化的划分,不是通过在这一区域使用这一形式体系,在另一区域使用另一形式体系的决定预先作出判断的。

(正像我将在后面强调的,我也认为,允许使用各种形式体系,在方法论上是很重要的:一个特定的子区域常常可以用

某种特异的方式简洁地加以表示。然而这并不造成矛盾：因为我们同时也主张，这种特异的形式体系能够系统地还原为基本的形式体系，因而将被看作“语义的结晶”。这是重要的，因为虽然一种表述方式的计算特性可能关键取决于这种特异形式体系的用法，但是必须有一个共同的表述框架系，在这框架系内，任何一个表述条目的意义内容都能够与别的任何条目的意义内容联系起来。）

我相信，可以在一个合理的时间尺度上构造出具有这些性质的朴素物理学的形式化。如此乐观的原因将在后面作出解释。然而，把这一方案同另一些可能与此混淆的方案明确区分开来是很重要的，因为其中有些方案的可操作程度看来要低得多。

3. 本方案不主张什么

(a)本方案无意编写一个计算机程序，使之能够在某种意义上“使用”这一形式体系，例如问题求解程序，或是以表述为目的的自然语言理解系统。我的兴趣所在始终是对其作出论证。（这些程序令人印象深刻，并且已令人满意地实际制造出某些有效的东西，如建立铁路模型；同时，某些学生也可因此而获得哲学博士学位。）然而程序会把注意力引向歧途。事实上，我认为它们还造成许多更危险的影响。很容易错误地作出结论：因为有一个（在某种意义上）有效的程序，所以它对知识的表述必然或多或少地（在某种意义上）是正确的。令人遗憾的是，为了使程序在合理的空间中或是合理的时间中有效，而必须采取的少量的折衷和简化，往往会使这一表述方

式甚至还不如本来那样令人满意。

这并非说，在构造本方案的形式化时，应当忽略计算问题。例如，普通常识性推理的派生长度的问题就很重要，同时我们的“稠密性”概念也对例如存储和恢复策略造成直接的计算后果。但是构造“证明”程序，似乎并没有服务于真正有用的目的(麦克德莫特也有类似看法，McDermott 1977)。

我之所以强调这一点，是因为在 AI 中有一种流行的看法：一项不能在某种有效的程序中产生立竿见影效果的研究，是没有多少用处的，或至少很值得怀疑。可能部分地由于这种看法，以致在表述方向上没有十足认真地尽力，同时产生出许多在不足道的小范围中工作良好(有时也很糟)的程序和技术，但这些程序和技术完全受到尺度因子的限制，因而，对于如何认识实际的复杂世界，什么也没有提供。(回溯式搜索和通过增加及删节表来表述动作的 STRIPS 方法，是两个有力的例子。我推测，产生式系统可作为另一个例子。)

理想的情况是：假定一个专门的推理机制，同时在更高层次上增加进一步的信息，用来“控制”该机制所完成的推理，这样做，原则上就应该能够从形式化中得到一个有效的程序(Hayes 1973; Kowalski 1977; Pratt 1977; Bundy 1978)。这样看起来，可以把形式化看作种种推理能力的“内核”，在执行特定任务的任何时刻，必须对这些能力的恰当展开作出进一步的详细说明。(然而由于特异表述方式具有特别合乎需要的计算特性，对这种理想的状况无疑还要作出修正。)

把实施细节放在次要地位的决策，其实蕴含着这样的主张：大型形式化的表述内容，可以与实施方式的决策十分清楚地划分开来。虽然我相信这一点是切实可信的，但它绝不是

显而易见的。

(b) 本方案**无意**创造一种新的、能把所有知识都记录下来的形式体系或语言。事实上,我(正像我的朋友们会猜到的那样)认为:对于表述来说,一阶逻辑是一个适用的基本工具。下面就来阐明这一点。

我并不想为通常的一阶逻辑**句法**作什么特殊的辩护。我个人认为它是可以接受的,但是这不排除有的人想用 KRL、语义网络、这种或那种"奇特的"语义网络,或是已有的什么东西把句法全部写出来。重要的问题是要**知道它的含义是什么**,即形式体系所具有的清楚**解释**(我有意避免用"s* m* nt* cs"这个词)。在解释层次上,这几方面之间没有什么可供选择的,而且它们大都严格地弱于谓词演算,谓词演算具有一个清晰、明确的模型理论,和一个易于理解的证明论,这也是它的优点(Hayes 1977,1978a)。

其次,我还要强调,特异标记法对于特异分支理论有时可能是有用的。例如,在概述流体的公理理论时(Hayes 1978b),我发现,把流体的可能物理状态在本质上看作有限状态机器的状态,是很有用处的。这样就把一些冗长而臃肿的一阶公理概括为一个简洁的图形。它的**含义**与这些公理一样:一阶事实上仍是参照语言。还有一些例子,如"可计值"的谓词和函数,有时出现在定理证明程序中,例如术语"2+3"被**计值**为常数"5",对算术式来说,没有可供使用的公理。但是它始终可以[①] 看作一种在计算上有效的表述方式,与"2+3=5",

① 这个提法忽略了一个有技术难点的棘手区域,然而这些难点不适合在这里讨论。

"2 + 2 = 4"之类公理的(无穷)集合表述了相同的**意义**。

第三,从一阶逻辑的实际情况看,它肯定是不够丰富而需要加以扩展的。我已经找到两种在我看来是必要的扩展方式:嵌入语,可使形式体系描述它自身的公式;一种非单值的司寇伦函数,类似于希尔伯特的符号 ε。缺省的概念也许可以作为另一种方式(虽然到目前为止我还没有感到特别需要这个概念)。我希望在使用形式体系遇到困难时,这种扩展会自然而然地出现。同时,我认为,想要事先预见这些困难的做法是危险的,所以我不打算这样做。

4. 公理–概念图:簇和稠密性

我们假定朴素物理学的形式化是存在的(我的看法是,它其实就存在于我的头脑中),并准备对它的结构加以分析。它主要是由大量断言组成的,其中包括大量(非逻辑)符号:关系符号,函数和常数符号(**或者是**:框式标题,栏目名称,等等;**或者是**:节点和弧形标号,等等。以后,不再赘述这些显然等价的东西,而假定读者熟知它们)。我们采用一个中性的词,把这些符号称为**标志**。

标志的**意义**是由形式化的结构,即由诸断言之间的推理联结的模式定义的。这个结构可能非常复杂,但是我们可以采用反映本质的定性方法来对待它,从而达到某种基本认识。

我们规定,形式化是**稠密的**,如果对于每一标志,都存在着许多包含该标志的公理。一种稠密的形式化在分散的概念之间形成许多联系,这些概念是用形式化中的标志来表示的。

稠密性显然有程度之分。因此,非稠密的形式化(**稀疏的**形式化)是不够理想的,因为它们不能足够准确地固定所含标志的意义。如果一个形式化规定(积木世界中)"在上方"关系仅仅就是:它是"位于其上"的可递闭合;并且,对"位于其上"的规定仅仅就是当一块积木没有任何东西"位于其上"时,就可以把它拣起来;而对"拣起来"的规定仅仅就是在**拣起来**之后就**持有**积木,等等(以此类推)——如果这就是形式化用这些标志规定的**全部**内容,那么其实很难说"位于其上"、"在上方"之类的概念已在形式化中得到表述。因为在我们的头脑中,这些概念还同其他许多概念相联结。如果一个东西在另一个东西上方,会出现各种结果。也许前一个的支撑垮了,落到后一个之上,于是两个物体的相对状况又有许多结果:顶上的那个可能为底下那个提供遮掩;如果底下的支撑住顶上的,那么底下一个就会由于顶上那个的重量而产生应变;如此等等。我们应当尽量捕获这种概念联系的**丰富性**。

〔值得强调的是,这里采纳的关于意义的观点,完全不同于那种认为形式化中的标志本质上是自然语言的词的观点(Wilks 1977)。根据后者的看法,通过**认可**,标志**确实**代表意向中的概念,即它们是"语义原素",可由它们构成其他所有的意义。此点稍后再论。〕

任何一种形式化,若想达到我们人类概念体系所拥有的丰富性,就必须是稠密的。当然,稠密性不是成功的**充分**条件,因为完全可以虚构一些全然无用的任意高稠密性的公理化形式。

最好来看一看形式化的简化模型。设想有一个图,它的节点是形式化的标志,它的弧形连线代表公理:一个弧连接两

个节点，如果弧所对应的公理包含这两个标志的话。〔严格地说，这必然是一个多重连接图形（Landin 1970），因为公理很可能包含多于两个的标志。然而，我们只是示意地利用这一思想。总之，这一技术还不够完善。〕我们称之为公理-概念（a-c）图。如果 a-c 图联结充分，这个形式化就是稠密的；如果图形稀疏，形式化也是稀疏的。然而，我们不能指望稠密性是均匀的：某些概念之间联系多，就出现了较为稠密的概念簇，而这些概念与形式化其余部分的联系则欠紧密。

在创立朴素物理学时，确认这些簇，既是最重要的，也是最困难的方法论任务之一。我认为过去已经出现过若干严重失误，例如，我现在倾向于认为，因果性不是一个簇：不存在一个有用的、或多或少自包含的因果性理论。“因果性”这个词的意思是，一些事情的发生同其他一些事情的发生及发生时间有关，而这取决于环境。例如，某物周围都是液体，所发生的事情多半就与在洁净干燥的环境中很不相同。然而，液体发生的情况，是**液体**簇的部分内容，而不是某个“条件-后果”理论的部分内容。这种错误是很难避免的，因为**可以**在任何地方进入一个大型概念结构。如果难以说出所提的概念在什么方面非常有用，这是一个出错的征兆（因为已在一个局部稀疏的地方，而不是在簇的某处进入了该图形）。但是这也可能是因为概念选择不当，缺乏想象力，或是别的什么原因。好在不难识别什么时候处在簇之中：断言本身会作出提示，比人们写出它们更加便捷。

超级簇也是一个有用的概念，这种簇与大量别的簇发生联系。我认为与三维形状和方向有关的概念集合，就是我们自己思想概念结构中的一个超级簇：诸如上，下，高，胖，宽，背

后，接触，位于，斜角，(表面的)棱，(体积的)表面，边，垂直，顶部，底部……这些概念。它们显然具有许多内在关系，因而它们构成一个簇。作为视知觉和空间运动的基础，它们也必然出现在任何一种概念框架系中，并起着重要作用。在描述组合体时，它们是至关重要的，在流体理论中也是如此(Hayes 1978b)。我还相信，在身体动作和事件的描述中也不例外(Hayes 1978c)，还有其他方面。

通过超级簇在各种其他簇中以这样的方式产生出来的事实，就可以将它们识别出来。其他可能作为超级簇的，有度量尺度理论(它会为各种任务提供像准确、模糊、功用这样一些概念)，时间度量理论，以及同**内部**、**外部**、容量和从一处到另一处的**穿道**有关的概念的集合。

对于这里提出的有关簇的术语不能太拘泥于字面：我无意指出我们的概念结构中有许多截然隔离的部分，它们能在与其他各部分完全隔离的情况下形成。同时，无论怎么说，公理系统的“a-c 图”模型本身在许多重要方面都是过分简化的。然而，我认为，存在着一些相互间有密切联结概念的集合，这一基本看法大体上是正确的，也是相当重要的。

5. A/C 比与还原论的形式化

下面我们来看一个还不十分成熟的形式化模型：公理与概念的**比率**(a/c 比)。对于一个稠密的公理化来说，a/c 是很大的。任何有价值的形式化的 a/c 比都大于 1，但是也有些有价值的形式化的 a/c 非常接近于 1。

我们看一看公理集合论。作为基础研究，这一思想是要建立一个**小型**公理论（例如策梅罗-弗朗克尔集合论，它的 $c=2$，即“ε”和“**集合**”，所以 $a=8$ 时，可得 $a/c=4$），通过该理论，我们能够**定义**大量数学概念（例如，整数可被定义为属于若干特殊种类中任一种类的集合；有理数是整数对的集合；实数是有理数无穷集的集合……），使得这些概念的期望特性（如整数的归纳原理，实线的连续性）**取决于**这些定义的结构，以及这一基本理论的公理。重要的是要认识到，被定义概念的这些特性是公理化的一些**定理**，而公理化是由集合论与概念的定义一起组成的。这些特性并不是固定被引入概念的意义所必需的公理假定本身，定义要尽可能完全地对概念加以固定，这样，所有其他东西就会随之而来。数学被还原为一系列集合论**引理**，或者至少想法是如此。（当我们这样考虑它时，看来几乎难以置信的是，这样一个大胆的计划竟已接近成功了。）

我想要强调的是，在形式体系中获取意义的这个方法与我提出的研究朴素物理学的公理方法是多么不同。在极端的情况下，集合论表现为还原论：我希望促成一个有丰富联结的形式化，并伴有各个假定之间的许多相互作用。的确，还原论方法会导致公理论，但这些理论是极为**松散的**。

我们来看看为一个**形式化**增加新概念的定义所产生的影响。a 和 c 同时增加 1。如果 a/c 是大的，这种做法会使其明显减小。的确，为了使这个比值（a/c）回到原来的值，不得不近似地增加许多公理。

这就强调了，直观上清楚的东西，即不能通过引进作为某种一般用法的概念，而使其获得意义的定义，很可能是错误的，它们会冲淡形式化。然而假定 a 和 c 都很小，比如说

a = 8, c = 2;增加一个定义后,a/c 就从 4 减小到 3;再增加一个,a/c 减小到 2.5;增加 1000 个定义,a/c 减小到 1.059。显然,随着形式化中定义数目的增加,a/c 渐近地趋向于 1。这种形式化的 a - c 图在中心有一个非常小的簇,被一团节点包围着,每个节点都呈辐射形地同较靠近中心的几个节点相连。这几乎就是可能存在的最稀疏的、并且包含所采用的概念标志的连接图。它的"形状"同稠密的公理论所具有的那种相互联结的、成簇的图形相差很大。

这种数学还原理论的存在,是一个值得重视的事实,如果能够为常识性推理找到一个类似的还原理论,一定会令人惊叹不已:一个小型概念集合,以及与这些概念相联结的公理,使所有其他概念(例如所有用英语词汇表示的概念)都能用这少数几个概念来定义。事实上,这是如此令人惊异,以致我感到可以肯定地断言这种小型理论根本不存在。可是,在 AI 的文献中,许多关于意义的形式表述的方法作了这样的假定。这就是"语义原素"的方法,在威尔克斯(Wilks 1975)和尚克(Schank 1975)的著作中就有这种例子。这里,小型标志集合中的元(威尔克斯的约 90 个,尚克的是对于某个 n 是 14 + n 个,n 是未知的)被当作**原素**。这样,一个英语词的**意义**就成为由这些原素标志运用某些形式工具建造的某种形式表达(在尚克那里,一般是用图形表示的,但这不是问题的实质)。用我们的话来说,形式化主要是由定义组成的:它的 a/c 比趋向于 1,正如公理集合论一样。在尚克和他学生的著作中,可以清楚地看到,"核心"理论的公理结构,意在起到集合公理在形成集合论时的那种中心组织作用(参阅里格尔的与 14 个作为原素的动作标志有关的"推理分子")。这就是说,非原素概

念(如“买”或“给”)的期望特性来自它们的定义,以及由核心理论赋予原素的意义。在威尔克斯的著作中,似乎根本不存在任何核心形式化:我们只看到标志表和假定这些标志所表示的概念的简要描述(参阅 Wilks 1977)。认为一个形式符号具有它所在形式化结构所规定的意义以外的某种意义,也就是说它具有某种固有的意义,这会造成一种十分令人遗憾的错误。然而,威尔克斯的观点是,他的语义原素事实上是词,如英语单词,而不仅仅是形式标志,这样就巧妙地避免了这个缺陷,并解释了为什么不必给出包含它们的任何形式化。我承认还存在着一个难题,这就是他的程序怎么知道这些词的意义是什么。

这个通往意义的还原论语义原素法,本质上必然是低精确性、低稠密性的表述。这种表述方式也有其用途,例如,它们可能适合于信息恢复或机器翻译这些应用,而且在它们确实行之有效时,具有某些非常有用的计算特性。但是有些时候,我们不得不面对表述世界的详尽知识的问题,这就要求放弃对于标志意义的“定义”观点。正如威尔克斯所说,“根本不能指望一个建立在原素上的表述方式能通过其结构来区别榔头、槌和斧……的意义”。情况也许并非如此,朴素物理学应该能做到这一点。

6. 意义、模型理论和精确性

有人或许会问:如果标志的意义不是由定义说明的,那么它们是如何被说明的呢? 从一定意义上说,这个问题是没

有答案的。我们无法指着一个特定结构说,**这**就是某个标志的意义。我们只能说,标志在如下程度上**意指**一个概念:**作为整体的形式化**能够使充分多的、在其结论中包含该标志的推理,亦即言及此概念的推理得以完成。如果形式化有一个恰当的模型理论,意义的操作性定义就能用外延法重新规定:在**作为整体的形式化**的每一可能的模型中,如果该标志代表一个实体,而该实体在模型所示出的可能的事物状态中,被人们认为对概念作出了令人满意的例示,那么,这个标志就意指这一概念。

要做到这一点,就必须有可能把模型看作"事物状态"。既然我们要对具有物理实在的常识世界加以形式化,这就意味着,对我们来说形式化模型必须可被识别为物理实在的摹本,因为在它里面,我们感兴趣的概念能够得到识别。

对一阶公理化而言一个模型就是一个集合——存在于用该模型表述的"事物状态"之中的实体的集合,同时,它也是从公理化标志到这个集合、以及覆盖这一集合的关系和函数的多个集合的特定映射。在初等逻辑教科书中,这通常是以相当形式的、数学的方式表现的,这一事实有可能导致一个奇怪的、却十分流行的错觉:一阶模型只不过是世界的另一种形式描述,正像以它为模型的公理化一样;而塔尔斯基的真值递归是一种从后者到前者的变换,从一个形式系统到另一形式系统的变换(例如 Wilks 1977)。这是非常错误的。从一开始,公理化和它的模型之间(或者双向地,在模型和它的正确的公理化集合之间)的关系就全然不同于一个变换。例如,它是多与多的对应,而不是一一对应。此外,它具有称为伽罗瓦联结的

代数特征，这种联结可以粗略地说成是，随着公理化规模的增长（即增加公理），模型集合，即可能的事物状态，在规模上有所减小。一个大型复杂的公理化完全可能具有小型简单的模型，反之亦然。特别是，一个模型常常会自动地变得复杂起来（例如将公理化中根本没有言及的一些实体包括进来）。但是这种思维方式更深层的错误是混淆了模型的**形式描述**和**实际模型**，这种形式描述可以在那些形成元逻辑理论的数学方法的教科书中找到。这就好像把结构工程教科书中对悉尼港大桥的数学描述同实际的桥梁混为一谈。塔尔斯基的模型实际上可以**作为**一个实体。如果我们有一个关于三个积木的积木世界的公理化，三个积木分别是“A”、“B”、“C”（这是在公理化中用于言及积木的标志），同时，如果我面前有一张（真的、物质的）桌子，在它上面有三个（真的、物质的）木块，那么这三个积木的集合就可以**作为**公理化模型的实体集合（当然是在这样的前提下：我能够不断地把公理化的关系和函数解释为施于木块的物理操作，或者不管怎样，只要在解释过程中根据公理化对木块作出的论断**在实际中**是正确的）。在塔尔斯基的一阶逻辑模型理论中，没有任何东西事先阻止真实世界成为公理系统的模型。

另一方面，也确实有许多公理化模型不包含实在的物理对象，但是在这些模型中，标志指称的可能是整数或其他符号。事实上，任何一阶公理化，只要它具有任何一个模型（也就是说它是无矛盾的），那么它就具有一个只有符号存在的模型——这是“赫伯兰德解释”，在这解释中，我们令标志指称它们自己。这样，塔尔斯基的模型理论并不保证公理化同任何特定的世界“有关”。常常会出现一致的看法：只有形式化符

号本身才是存在的。[1] 这或许可以称为"唯我论"解释:不承认外部世界的存在,同时却有一个关于它的周详理论。

由此可知任何公理化都有多个(常常是无穷多个)模型,我们不能仅仅举出其中一个与真实有些相似的模型,就证明该公理化是一个恰当的描述。因为该公理化可能还有一个比它简单得多的模型。如果公理化有这样一个比较简单的模型,那么出现在公理化中的标志的含义,就不多于它在这一简单模型中的含义。这正是我所说的"精确性"。比如说,积木世界的低精度的形式化所接纳的模型比预期的要简单得多。例如这样的模型:其中的"积木"是整数,"在上方"意味着大于,等等;或者是这种模型:其中的"积木"是离散的二维空间中的一些点,或者其他什么模型。一个积木世界的恰当的形式化,即一个高精度的形式化,应该是这样的:它的任一模型都必须有一个本质上是三维的结构。因此,SHRDLU 的积木世界公理化(Winograd 1972)采用了三维笛卡尔坐标。(找到一个有用的、但定量性较少的描述三维结构的方式,是很有意义的。例如,一个刚性连件需要三个接触点:两条腿的凳子站不稳。)

了解形式化精确性的一个好办法,是看它的**最简**模型与**预期**模型的相似程度如何。我认为,这极好地证明了表述性语言具有一个模型理论,因为它给了我们一个检验表述精确性的方法。人们很容易滑向这种看法:既然在一个十分简单的模型中,就形式化所及的全部内容而言,事实上标志有可能

[1] 虽然通常的一阶逻辑的形式化不能表达这种信念,但是至少在一个外延中是可以表达的,它的否定形式——存在着某些不是符号的东西,同样能得到表达。

完全表示某种更为基本的东西，那么该形式化就已经捕获了这一概念（例如，由于已经用了一个听上去可信的标志来代表这一概念）。

（这一判据表明，为了使最简模型具有一定的复杂性，我们应当对那些能够在形式化中使用的表述性语言的特征予以注意。在一阶逻辑场合，它们包括函数[运用函数就意味着，对于任一适合的自变量，都有一个函数值存在]、明确的存在性论断[特别是各种各样的“概括公理”——详述见后]、等式的应用、高等分类逻辑的应用[用 AI 的行话来说，是存在着与量词有关的 isa 层级体系，虽然它可以比简单的层级体系复杂得多，参考海斯(Hayes 1971)文中使用的分类逻辑]。因此，我们或可期待着这些特征在朴素物理学的发展中起到重要作用。）

根据对标志的意义的这种解释，它取决于标志为其组成部分的形式**化整体**。这样，从原则上讲，任一部分形式化的变化，都能改变它的其余每一部分的意义。同时，我认为这基本上是正确的：标志的意义，从内省角度看，就是得知一个新的事实或掌握一个新的概念，这很容易对人们理解其他概念的意义的方式产生深远的影响。这同时也意味着，头脑中有着不同形式化的人，可能以不同方式理解同一标志。“水”代表的意义对你我而言不会分毫不差，可能会有这种情况：看到一种物质和一组场景后，我把它叫做水，而你却不（例如，如果你从未见到过不透明的水，你也许否认我桌上玻璃杯中这种不透明液体是水）。然而，**我们俩也许都是正确的**，因为我们关于“水”的理论可能不完全等同。即使我们对水的认识（在对水的直接感受中，即在作出所有实际上包含标志“水”的论断

时)是等同的,情况也可能如此。我们每个人也许同意另一个人对水的每一看法,但是我们的概念也许有细微的差别。这种差别可能存在于一些我们对之有不同理解的相关概念中(像粘性、可饮用性)。甚至要准确说出我们存在分歧的是哪些标志,也是不可能的:只能说我们关于这些标志的理论是有所不同的。因而对一个标志来说,不存在单一的"意义",或者至少若假定它的存在,就等于假定人们的认知形式化在结构上是等同的。如果在思想中不对这一点保持清醒,许多混乱就会接踵而来。例如,威尔克斯曾经论证(Anderson et al. 1972)既然有一些人从未见过冰,那么水结冻这一事实就不能作为"水"的意义的一部分,因为假如这样做的话,就不得不说那些人不理解"水"的意义,这显然是荒谬的。这个词的含义对一些人来说显然比另一些人要多,通过对这一事实的观察,我们可以避开这段曲折的推理过程,而对于的确知道冰的人来说,冰是结冻的水这个事实可以很自然地成为"水"的含义的一部分。

然而,要使交流成为可能,人们的认知结构显然应当是相似的,作为一个行之有效的前提,在形成朴素物理学时,我们将采用这一假定。我们之所以首先把朴素物理学选为研究的对象,一个重要原因就是,在这一领域中,人与人之间的一致程度似乎比许多领域更大些。

看来,形式化中还要有一个"距离"的概念,使一个标志离变化越远,变化对标志意义的影响就越小。我尚不清楚,这个提示性的直觉看法是否能够成立,又如何能够成立。可以认为这一距离等同于公理-概念图中的最短路径距离,虽然这不一定合适,因为它忽略了公理的结构,但是,这是目前我能采

取的最好做法了。

由于这一距离弱化效应，看来首先多少有些独立地从簇入手，是一个合理的策略，因为簇的结构对簇内标志的意义的约束，比起同其他簇的联系的约束来，是更加严格的。因而相当自由地引入一些确切地出现在别的某个簇中的概念，假定它们的意义相当严格地由，或将由，那个簇加以说明，看来是合理的。例如在考虑液体时，我就要能够谈论体积形状，即假定——我现在认为是合理假定——形状簇将为我说明这些液体。毫无疑问，它们出现在液体簇中，**的确**改变了原来的意义：假如我们从未见过一个巨大的静止水域，就很难得出水平面的完整概念。但是，假定一个**完全自成一体的**形状理论，看来仍是合理的。

这里，说到底我是在主张，虽然对意义作"定义"的观点是错误的，但是为了取得进展，我们可以——甚至是必须——把它看作**仿佛**是正确的。它是一个理想的方法论脚手架。

关于意义的模型理论观点还有最后一点看法。正如我已经说过的，任何相容的一阶公理化都有一个仅含符号的模型。尽管最简单的这种模型可能是相当复杂的，但是人们或许感到，如果它所包含的全部是**符号**，可能很难认为它与真正的物理世界有相像之处，即使在某种意义上，它们在"抽象"结构中有相似之处。

为了对这一不同观点作出回答，必须谈一谈身体和感觉的输入。让我们设想朴素物理学形式化有一个带感觉器官的（有形的）身体。形式化得以附着于物理世界的方式，是运用"实在论"的观点来看待由它的知觉为它提供的数据。这样，朴素物理学就应该部分地成为一个**表象**理论。由于视觉方面

的研究成果,这一理论目前正处于发展之中(特别是玛尔和霍恩的工作,他们有意识地尝试把表象和形体结构联系起来)。虽然一个详尽的高保真的积木块理论很可能只不过是一个梦想,例如积木标志在某些模型中也许表示(比如说)整数,但是如果这理论也说明了(在过分简化的情况下)当一个人以**此种**取向和**此种**照明条件凝视一个有**此种**表面的积木时,他就会看到**此种**映像,那么上述说法就将是不正确的。因为整数或符号并不是人们可以凝视的那种东西。(即使从某种特殊的意义上说有这种可能,也肯定与看一块砖不一样。)或许会有这种反对意见:但是你们**假定**的"凝视"的标志必然是指这样做的身体动作,这是用未经证明的假定来辩解,比如说,如果一个整数模型也满足公理化,对这模型中的这些标志作出解释也是可能的。是的,我正是要用未经证明的假定来辩解:我假定"运动标志"——描述身体运动的符号——**直接**同身体发生联系。它们构成了与(实际的)身体有着十分特殊关系的**身体映像**[①](我把它想象成类似于图形数据结构和屏幕上的物理画面之间的那种关系)。

因此这个假定的意思是,朴素物理学可以与真实的物理世界"相联结",因为它有一个配备着感觉器官的物质的身体:**凝视**(或**感觉**、**推动**等)的概念**本质上**是物质的,它之所以具有这个特点,是由于它在身体感觉运动系统中有固定的解释。正是由于这种附着于公理化整体上的公理化中某些标志所作的固定的、物理的解释,才使得公理化必然包含着真实的、物理的实体和关系。因而在某种意义上,人们可能预期朴素物

---

① 我十分感激 S·魏尔向我介绍这一概念。

理学的大部分都“靠近”视觉表象(或触觉、嗅觉、听觉)的概念簇(一个或多个),所以朴素物理学的任一部分都不会与感觉证据相距很远。然而,当精心设计的理论融入真实的物理学时,因为物理学包含着一些是由推导而来的远离上述证据的概念,常常存在一种有些不安的普遍情绪。例如在科学哲学的著作中,人们会听到谈论**理论性实体**(电子就是一个很好的例子)。在我看来,这种议论往往缺乏一种足够有力的认识方式,根据这种方式,即使像“一段木头”或“是湿的”这样的日常而平凡的思想都是同样理论性的构造物,尽管它们是在一个不同的、更为朴素的关于这个世界的理论中。

这一点给朴素物理学的启示是,应当随时抓住机会,把概念同感觉或感觉运动概念联系起来。例如,在研究液体时,我发现空间运动概念非常有用,在其他领域内,也同样有明显的功用。这可以直接同视觉理论建立联系。如果你朝一个存在着运动的空间看去,你就能**看到**这个运动。厄尔曼(Ullman 1977)十分详细地解释了怎样看到它。另一方面,我认为我们的内省常识世界的组织的丰富性有很多来自我们对做诸如推、拉、举**这些事情时感觉如何**的知识,即最终来自我们身体关节和肌肉中的本体感受器。我并不乐观地认为我们能在不远的将来用形式化获取这种丰富性。(这需要构造一个适当的、配备有必要感觉的有形身体。)但是我想,通过留意那些“附着”在身体运动概念上的概念,我们可以对形式化加以注释——这可能是一种有用的、同时也是有趣的训练。

这整个身心关系领域都是值得深入研究的,我相信 AI 概念能够为澄清这一关系而作出贡献。

7. 彻底性和闭合性

看来要得到高 a/c 比的一个方法也许是减小 c,并对某些新概念作大量说明。假如能够做到这一点,那的确是大有裨益和令人鼓舞的,要是我们能够找到一些小的、自包含的概念组,能够在完全隔离的情况下对它们进行形式化,并达到合理的精确程度的话。

然而这样的概念组看来并不多。(几何形状可以算作一个。)人们发现的典型情况是:刚刚选择了一些概念供开始用,马上就需要为另一些不期而至的概念引入许多标志,为了固定它们的意义,还需要引进更多的概念,于是标志激增,到了失控的程度。如果你把这看成对我们的概念结构的 a-c 图的探究,当然就不会对这种现象感到吃惊(特别是,如果我们假定这个图形是非常稠密的,而这是必需假定的)。处在当前的簇中,在识别进入另一些簇的路径时,我们需要有方向意识。然而,即使有这种意识(仅由经验形成的意识),必要概念的激增也几乎从一开始就是令人震惊的。

但是这种激增最终必然减缓下来:因为形式化是有限的。"彻底性"条件的中心思想是:**一直发展,直到它减缓下来**,直到发现概念集合已经自行闭合。这样,你想通过形式化说明的所有事情,都能够用已经引进的标志来说明。在图形类比中,则直到我们**跨越**整个图形为止,而只需要增加新的弧来填充图形,使它的稠密性足供捕获标志的意义。

建造过玩具世界公理化的人,都不会对闭合性这一思想

感到陌生。你突然发现,周围有足够的概念,使你可以说所有这些已"足够"了,所谓足够,就是说它们能使那些一直在你心里的推理得以实现。闭合性可以在非常小的形式化中得到,但是如果形式化是闭合的,**同时**有高精确性(因而有高稠密性),那么我相信它必然也是彻底的,因为它的范围必然涵盖了常识性推理的**全部**重要概念。这相当于说 a-c 图被相当强地联结起来,不存在任何真正隔离的子图形。

彻底性和精确性之间的这种相互联系有个程度问题。为了获取更高的精确性,需要有更大的彻底性。为了**真正**获取"在上方"的概念,仍停留在朴素物理学内部很可能是不够的,例如,为了论及人际间的地位高低,不得不涉及多种多样的类比。(法官的职位高升了;天堂在上,地狱在下;表示自己的谦虚和卑微;等等。)只有一个非常**广泛**的理论才能(经由模型理论的伽罗瓦联结)集中力量把"在上方"这一标志的意义约束得使它与我们的"这个"概念完全相配合。(设想在一个世界中,对"地位"的类比是颠倒的,以致处于某些人**下方**,就是支配和/或胜过他们。这将是一种可能的朴素物理学模型,但不属于更大范围的常识理论,而且这个世界与我们的世界全然不同。)如果形式化缺乏广泛性,就必须深入发展,而且必需深入到变得稠密,所以一个稠密的形式化必然是深入的和广泛的。

在这意义上,簇恰好是局部闭合的。一个簇包含一组在**某种**程度上相互闭合的概念,因为虽然也需要其他概念,但是在簇的内部,有大量东西可以看作是属于簇本身的概念的。因此,簇的形成也有一个程度问题,是由精确性和详尽水平决定的。

在与常识的其他部分相隔离的情况下，处理朴素物理学的整个过程，是以以下观点为基础的：存在着一个详尽水平，在这一水平上朴素物理学在一个更大的概念结构中构成了一个相当闭合的簇，这是一个丰富的、但可以驾驭的详尽水平，我认为它代表的彻底性同以前达到的相比，要高出一个数量级，但是不可能高出很多，比如说十个数量级。

8. 某些可能的簇及其概念

这一节中，对有关概念簇的某些具体思想作一概述。这只是一个梗概，比较粗略，更充分的说明最终会在别处作出，我不打算开一个详尽无遗的清单。

很可能有这种情况：这些簇中有很多并不是真正意义上的簇。例如，经过更加严密的研究，它们可能分为若干小部分；或者，可能揭示出一些新的连接关系，使得簇的界限模糊不清。然而我认为，作为探究的起点，它们是很适合的。

度量的尺度

我们应当有能力表达像大小、范围、（液体或粉剂的）多少、重量、粘度等这样一些量，它们看来是物体的特性。但是我们能用例如磅或千克来测定重量，所以就必须引进各种各样度量相同物理量的度量尺度的概念。我们可以拥有各种各样的从物体到（比如说）有理数的函数，将其称为磅重和千克重等等，但这是不方便的，不自然的，也不足以支撑一个非常

稠密的公理集合。我认为,应当引进一个**重量**(**大小、多少**)的"抽象空间"概念,这样,**重量**就是一个从实物到重量的函数,而磅(及其他)就是从有理数到重量的函数,于是可以写出:

重量(某人) = 磅(150.32) = 千克(68.25)

我认为这些重量、大小等**度量空间**都有它们自己的理论。它们很可能具有一个容许空间的结构(Zeeman 1962),即它们有一个有限的"粒度"。它们是各种各样的近似、接近和"典型"度量(如大象的标准尺寸)的概念,以及不等式和其他相关事物的概念,同时我推测,这之中有很多是独立于被度量的特定量的。

有一个评论也许适合于这里的情况。人们常常认为,"常识"需要一种不同于一般的模糊逻辑。作为支持这一观点的例子,无例外地包含着模糊度量尺度或度量空间。我认为在这里或许有模糊性的一席之地,但这不是对模糊真值的论证。

形状、方向和维度

三维物理形状。尽管在机器人的操作语言方面作出了一些值得称道的研究(参阅 Bolles 1976),但是就我所知,对这个簇的研究并不是很多。它还同视觉论题有联系,因为这方面已有许多成果,从这里入手可能是个好办法。我原希望能在这里对形状作更多的说明,但实际上只能作一些松散的讨论。

对朴素物理学来说,垂直重力是生活中一个恒常的事实,所以垂直维度和水平维度应当区别对待:"高"和"长"是不同的概念。一个物体在背靠一个刚性表面(比如墙)时和独立站

立时，对它形状的描述常常是不同的(宽和长；或是离墙的深度和沿墙的宽度或长度：如果一个物体被看作是靠墙而放，就是宽，如果被看作是沿墙伸展，就是长)。详尽的研究尚未作出，我怀疑这些不同的概念集合，来自于各种坐标系之间的协调。例如，一堵墙以沿其法线的半轴定义着一个自然坐标系。

形状的一个重要方面是面与立体和棱与面的关系。那些适用于许多专门场合的不同名称，表明了该簇的丰富性：顶、底、边、缘、棱、唇、前、后、廓、端。《罗热类属词典》(Roger's thesaurus，第二类，§2)提供了几百个这样的词。此外，在方向变化时，特别是与垂直重力有关时，它们不是不变的。这些边界概念在描述空间形状时也是至关重要的，而且还是同调论和微分几何的基础。

内部和外部

我们来看如下的概念集合：(内部)，(外部)，(房门、正门、窗、大门、入径、出径)，(墙壁、边界、容量)，(障碍物、屏障)，(过道、穿道)。

我认为这些词提示了一簇相关概念，这些概念是朴素物理学的重要基础。这个簇关系到把三维空间划分成有物理边界的块，也关系到这些空间部分可以相互联系的方式，以及物体、人和液体怎样能从这一位置到达另一位置。

我之所以认为这个簇重要，有几个原因。一个原因仅仅是从内省角度感到它是重要的。另一个原因是，这些观念，特别是"穿道"的观念和那些可能因此出毛病的事物，似乎是民间传说和传奇文学中常常出现的话题，也是许多常用类比的

根据。又一个原因是,在观察其他簇特别是像液体和历程(见后)这样的簇时,这些观念已相当频繁地冒出来。还有一个原因,它们是某个重要的数学理论——同伦论的来源。但是,主要的原因还在于**包含性对于因果性的限定**。待在房间里的一个主要理由,是把自己同外边正在产生的因果影响隔离开来,或是防止屋内的事泄漏到外边去(如分别是:为了避雨,为了进行密谋)。熟练掌握何种障碍可有效地抵制何种影响,看来是一种非常有用的才能,这是能够解决"框架问题"所必需的。

把这些观念同形状观念作对比是很有意思的。这里,我们所说的空间是一个可以位于其中的地方:可以说是**运动空间**;而在描述形状时,空间是物质**占据的空间**。然而,许多概念在这两个领域中都是有用的。

历程:对发生过程的描述

由J·麦卡锡开创的、描述动作和变化的现代经典方法,采用了**状态**或**情境**的概念。它被看作是在给定瞬间为世界拍的一张快照,因此动作和事件就成了从一些状态到另一些状态的函数。观念的这一框架甚至被许多否定自己的形式体系中包含状态变量的人所采用,并且已经被特意编入数种 AI 程序语言。然而我现在认为,这是一个错误,至少是一个粗糙的过度简化。

我们来看以下例子(这是 R·伯斯塔尔多年前让我看的,但当时我并不懂):在纽约,有两个人约定一个星期后在伦敦见面,然后他们分头出发:一个去爱丁堡,一个去旧金山。他们每个人独自度过了丰富多彩的一个星期,然后,按照事先的

安排，如期相会。为了用情境来描述这事，我们必须说明，他们之中每一个人在另一个人发生的每一件事之后，发生了什么事情，因为每一情境既然在概念上是整个世界的一种状态，就包括了他们两人。

我们所需的是具有限定空间范围的事件状态的概念。我所谓的**历程**是这样的对象，即相联结的时空块，一般被界定在其中有“某些事情发生”的四维坐标中（这里也包括无事情发生的特殊情况）。

历程的三维空间截面，是特定时刻的一个地点，即该地点的状态。地点可大可小，可相互嵌套：房间、旅馆、街道（可以理解为全部面向街道的建筑物组成的内部空间）、城市，都可以是地点。例如，典型的历程是在某天下午 1 时到 4 时之间一个特定房间的内部。从概念上讲，空间是由地点构成的，而时空是由参差交错的历程构成的。我们也可以把历程看作一个过程（的发生）的**展开**。

任何一个明确定义的物体或空间块，都可以乘以（在代数直积意义上的）时间间隔，平凡地展开为历程。但是也有一些历程不那么简单，如弹道，它在时空中是“倾斜”的。

能够以物体或空间块以及历程的**形状**为参照，是十分有用的。例如一个下落的（如从罐内倒出的）水柱定义了一个历程，它的形状是一个垂直柱（准确地说，这不是柱，而是空间旋转体——译者注）。在这一历程同液体正在进入的别的历程的联系中，这个柱的顶部和底部是相当重要的。

历程可以通过各种方式相互发生联系。有**相邻关系**：既有空间的（例如，水柱落到桌面上时，是垂直在上并相接触），也有时间的（例如，接触开关，随即启动发动机，是立即跟随），

以及混合的(如飞行器之间的碰撞,是两个轨道的相交)。在地点之间,因而也在历程之间,存在着形状特性和相对位置关系。有相似地从地点和物体那里继承来的空间包含关系,也有时间包含关系("当")。还有历程与各种整体坐标系之间的既是时间也是空间的关系,我们可以称之为历程的**编址**。有很多可行的编址系统,并非都是米制的坐标框架,例如一座大楼中的房间编号系统。所有这些定义出了可以称为朴素时空几何学的东西。

并非每一空间区域都可以作为明确定义的地点,也不是每一时空区域都可以看作历程。"棱"的定义必须由一些"自然"边界作出。可作为自然边界的东西是(审慎地)无限制的,但是物理屏障是显而易见的例子,比如房间的墙。

既然地点可以嵌套(的确,或许每一地点都在某个另一地点**内部**),那么每一事件也包含在许多(也许是无穷多的)历程之中。但是对于每一类型的事件,都有一个严格包含它的最小历程,即这类事件的因果影响无法通过的那些屏障在空间上界定的最小历程。〔**粗略地说**,对于"(事件)是在哪里发生的?"的问题,这个地点是自然的答案(可作为答案的有:在桌子前,在居室内,在那间屋子里,在伦敦)〕。这一观念的重要性前已提及:这样的屏障限制了事件的因果影响必须被追溯的范围,因而使预见变得更加容易。根据对屏障几何形状的**静态**描述,可以预见各种事件只能影响为数不多的历程。例如,在关闭的房间里发生的很不平常的事件,只有有限的一类(大爆炸,洪水,大火)才能直接对房间以外的历程产生影响。

同经典的情境/动作本体论相比,历程给我们带来的表达能力和预见能力的提高,还表现在其他若干重要方面,由于需

要的篇幅太长，不在此细述。更充分的说明尚在准备之中。

能量与作用力

在作预测时，区分“直接发生”的事件（如下落）和需要某种作用力或能量消耗的事件（如飞石穿越空气），看来是十分关键的。显然，这一观点是，如果在已知历程中没有施加作用力，那么后一类事件就不会发生。

这种区分违反了能量守恒定律，但我认为对朴素物理学来说，这样做是完全正确的（也许我们只能说，“作用力”的直觉概念并不确切地对应于物理学中“作功”的概念）。在许多日常情境中，能量消耗几乎没有什么明显的结果（例如向砖头里钉钉子）。

我无法肯定关于作用力的概念还能说些什么，也许作用力来源的容量是有限的（它们耗尽了或疲劳了）。也可能与其说它是一个簇，不如说它是一个把其他许多簇松散连接起来的概念。

执行者可以作为能量的来源，但是这两个概念是有区别的，因为有一些动作不需要能量（如讲话），而一些能源不具有意志力。然而，在朴素物理学中这两者也可能是等价的概念，只有运用像意志力这样的“心理”概念才能将它们区分开来。

组　合

许多固态物理体是以某种方式将部件组装到一起而构成的；而另一些则只是一块（某种）原料，如一段木头。有相

当多的概念与组合这一观念相联结，像这样一些概念：作为元件、作为部件〔例如人的手是人的一个部件，而不是元件，人的肝脏则是元件（尚有争论），因为在某种程度上它是可分离的〕、附着、组合和拆开、粘接、钉入、拧上等等……。还有一些概念与被组合部件相互间可作相对运动的方式有关（轴、滑轮、键槽、铰链），这些概念与形状和运动的空间几何概念相联结。还有一些概念与不同类型的材料的机械特性有关：刚度、硬度、柔性、易切割、易粘接等等。

支　撑

如果听其自然，物体（或液体）就会下落。为了制止它们下落，就必须把它们支撑起来。我想我们可以把所有支撑一个物体的方式列一个简表如下：

1. 将某物放置其下，托住它。当然，该物也要有支撑，以此类推。但是**地面**无需支撑：**它**是所有支撑关系的基础（由此推得，地面不是一个物体）。

2. 某物位于其上，它**挂在**该物上。

3. 某物紧靠它，它**附着于**该物上（参阅组合）。

4. 它**漂浮**在某些被包围的液体上。

5. 它**正在飞**，即不接触任何固体，以某种方式自我托起。这需要飞行着的物体要有很大的作用力，所以没有生命的、“被动的”物体不能飞（尽管风筝是个例外）。

在以上情况中，(1)是最安全的；在其他所有情况中，某一组成部分失效〔分别有：断裂（如绳索）、分离、泄漏、作用力停止——所有这些都是这种或那种历程〕，就意味着支撑结束，

因此下落历程会突然开始，而下落历程（常常）会造成危险的结局。这样，在“从下面支撑”的想法周围，存在着一个概念的微型簇：如高大建筑、塔、墙等这样一些概念，以及稳定性和可能失效的方式（翻倒、破碎、散开、滑动、下沉）。

物质和物理状态

有许多不同种类的物料：铁、水、木料、肉、石头、沙等等，它们以不同类型的**物理状态**存在着：固态、液态、粉末、糊状、胶体、泥状、纸状等等。每一种物料都有一个**常见**状态：铁是固态的，水是液态的，沙是粉末状的等等，但是这一点有时也会改变。例如，许多物料加热到一定程度，就会**熔化**（对有些东西来说这是**非常**之高的温度，也就是说，**实际上**它们不可能熔化，例如沙；还有一些在加热时会**燃烧**，例如木头或面粉）。如果让液体冷却到一定程度，它们就会凝固。如果你以足够的力量和决心来研磨固体，它们就会变成粉末，如此等等。没有显然的常规方法能将粉末变成固体（但是把它们打湿，成为糊状，再仔细烘干，往往行之有效）。

有时，对于以两种不同状态存在的同一物质，我们会得出两个概念。沙和岩石即是一例。我认为这是有其理由的：当物质（1）从一种状态到另一状态的转换是极端困难的时候；（2）两者都以自然形式存在的时候。（以铁锉屑为例，它满足（1），但不满足（2）。）

有些物质，任其自然存在，就会**分解**，即慢慢变成另一种（无用的）物质；或是**成熟**，即慢慢变为另一种（有用的）物质。生锈和潮湿腐烂都是分解的例子，制作奶酪则是成熟的例子。

每个不是组合的物理体必然是由某种物料构成,该物体的许多特性其实就是构成它的物质的特性(刚度、颜色——在未着色的情况下、硬度等等)。我认为把一个东西的这些特性同那些主要是与它的形状或结构相联结的特性区分开来,是十分重要的。某些物体主要由一种物质的特性确定(如一块铅),另一些物体则由别的特性确定(如一个建筑群)。有些特性,像重量,既同大小有关,也同材料有关,随着物体种类的不同,它所具有的内涵也不同:一块重的铅就是一块大的铅;而一块重的建筑用砖,必须由某种特殊密度的材料制成。固体物体必须由物理状态是固体的材料制成,因为只有固体才可能看作是有形状的。由此可知,如果把固态物体加热到制作它的材料的熔点,它肯定就不作为一个物体而存在了,因为所需的那种物质状态已不存在。我想这是对熔化所作的一个有说服力的解释。

要探索各种半固态物质物理状态之间可能存在的转变,烹调看来是一个理想的领域,另一个则是制造加工过程(模制、铸造、锻造)。测定总量和分量的各种方法也是一个有待探索的实用领域。以木头和金属为例,它们(基本上)是根据重量或体积批量出售的,但是在各种不同的系统里也根据物件的形状(条状、平面状或立体)进行零售。

力 和 运 动

朴素物理学是伽利略之前的物理学。我在11岁时,因为听到讲授牛顿"运动定律",在理智上受到的震动,至今记忆犹新:没有力作用其上,一个物体怎么能保持运动呢?读伽利

略的"关于两大世界体系的对话"(1632)时,感到很有趣,他从日常经验出发,令人信服地论证了牛顿第一定律必然成立。但是这要用大量小心翼翼的论证,还依赖于读者对平坦、光滑的平面和近乎完美的球体的体验。另一方面,每一个儿童都具有的非牛顿式的直觉却是:一粒由弹弓射出的石子,是按辐射方向向前,而不是沿切线方向飞出的。还能找出其他一些例子。

一个正在运动的物体只有五种可能的运动方式:它或是**下落**;或是被某物**拉**或**推**;或是**自己**运动向前,边运动边消耗作用力(因而它不可能是被动物体);或是**滑动**(包括在光滑平面上的滑动和沿光滑斜面向下的滑动);或是**滚动**,在这种情况下,它必须是可滚物或轮子,或者有可滚物或轮子。在后两种情况中,作用力停止后,运动可以持续一段时间。(我们可以称这种现象为**惯性**滑行:这是向伽利略作出的姿态。)上述说明未包括旋转和摆动这两种运动:就**位置变化**而言,可以说它们包含了所有运动方式。

我认为在形成运动概念时,实际上有两种不同的方式:作为位移,或是作为轨迹。位移运动需要作用力,力停止了,运动就停止:它的特点在于始终存在一个相对于位置的恒定的伺服控制,亦即它的概念是**位置变化**。轨迹运动具有惯性,在(因碰撞)而被迫停止之前,继续前进,其特点是**沿一路径**作平稳运动。它的启动、停止或改变方向都需要作用力,而维持运动则只需较小的力或完全不用力。例子有:抛射体,汽车,溜冰,跳跃。位移运动是希腊式的,轨迹运动是伽利略式的。像瞄准、碰撞、速度(作为一个度量空间)、加速度这样一些概念同后者有关;而去、来、躲、避、趋向、远离这样一些概念是同前

者有关的。位移和轨迹，这两者都是历程，但是前者基本上只是从它们的起点到它们的终点的转变，起点和终点都是物体的位置；而后者有一个确定的**形状**，例如，它们可以在时间上作外延，因而有瞄准的概念。下落、滑动、滚动和跳跃都是轨迹的例子。

力可以用各种方式传递。刚体可以传递推力，绳子可以传递拉力。卢格尔和邦迪（Luger and Bundy 1977）较为详细地考察过这个微型簇。

液　　体

液态物质提出一些特殊问题，因为与固体材料的块不同，液体的“块”一般不用作为特定液体块的方法来区别，而是用使其处在特定**地点**（如湖）或是与固体的某种特殊关系（在杯子内）的方法。在拙作（Hayes 1978b）中，我详述了解决这些问题的方法。

9. 某些结构式形式化技术

构造“启发方式恰当”的公理形式化（McCarthy and Hayes 1969）是一门艺术，就像编写好的程序是一门艺术一样。它尚未得到充分发展（诚然，研究朴素物理学的主要目的之一，就是要在这个相对来说还是处女地的领域中取得一些技巧），但是有一些独特之处正在崭露头角。

其一是**分类学**的重要性：也就是对一种事物或是一事物

的可能状态的各种型式或范畴列出有限的、穷举的表。我们已经看到的有:被支撑的方式,物理状态的类型,一种流体的可能状态(有六种:被包容,流动,喷洒,打湿,下落及飞扬)。对于每一情况,我们都有一组如下形式的公理:

$$\Phi \equiv \Phi_1(x) \vee \cdots \vee \Phi_n(x)$$

$$\Phi_1(x) \supset T_1$$

$$\cdot$$

$$\cdot$$

$$\cdot$$

$$\Phi_n(x) \supset T_n$$

这里,T_i 是理论,表示那些特定场合的具体情况。这种穷举列表在作出推理时可能是非常有用的,方法是对在视觉中广泛使用的图形相容性检验计算加以概括(参阅 Mackworth 1977)。从直觉上看,“当且仅当($\equiv$)”的意思是,如果除一个析取项之外的所有析取项都可以排除掉,那么余下的一个必然就是所指的情况。如果次级理论的集合(T_i)被恰当地结构化,那么它就可作为一项强有力的获取简短证明的技术。例如,在**支撑**情况中,我们可以迅速作出推理:一个被动的东西用绳子吊着(无其他支撑),绳子断了,那么它必定下落;因为紧靠其下没有东西(情况 1),也不存在使它漂浮于上的液体(情况 4),它也没有附着在任何东西上(情况 3),同时它又不会飞(情况 5),所以它必须由吊挂来支撑。于是(根据基本的**支撑**公理)它必然下落(亦即这一瞬间是下落历程的开端)。类似的论证方式也可以用来证明流到桌边的水将会下落(而不是比如说继续水平前进,或在桌边堆积起来)。

这种分类学具有句法形式的定义(根据 Φ_i 来定义 Φ),也

许是值得注意的，但是它们并没有起定义的作用，因为这个“被定义”的标志已经在公理化的别处出现过。

第二个独特之处与存在公理和概括公理有关。正如早些时候指出的，那些设定实体存在的公理，对于想要拥有非平凡模型的形式化来说，是关键性的。我们已经遇到的例子有：由物理边界(房间、茶杯内部)，或是由各种(米制的和非米制的)坐标系所定义的空间(地点)；当各种状态存在时，则有种种历程接踵而来，例如一个物体处于无支撑的状态时，必然随即发生下落。在这些情况中，当然还有别的情况中，有**概括公理**存在，它们确定了所需实体的存在，以及它与已有实体的关系〔墙壁**之间**或门**背后**的空间；在物体失去支撑的**瞬间后**(在该地点的下方)将发生下落；如此等等〕。

然而，我在这里要指出的是，这些都是**受限制的**概括公理。我们不能任意取出几块三维空间或四维时空，把它们看作是独立的，可以说它们只是用可描述方式与我们已确知的实体相联系的东西而已。我认为，这种在本体论承诺方面的选择，是常识性推理区别于“硬”科学或哲学推理的特点之一。常识性本体论是冗长的——所有种类的实体，具体的和抽象的(物体、材料、颜色、空间、时间、历程、事件，……)，在使用时缺乏哲学的严谨性，几乎不对基础的本体简单性作什么要求(例如与亚原子物理学甚或化学周期表作比较)；当然，它也是受到应有控制的，与古德曼(Goodman 1966)的唯名论有所不同，也与公理集合论或 λ 型演算的概括公理体系有所不同。这两种不同产生的影响，是得出一个具有远较以前丰富的结构式的实体集合：其个**数少于**这些“统一的”形式化，但是**种类**却多得多，同时它们之间的**关系类型**的集合也丰富得多。

还要指出一点：使用统一的米制坐标框架系，实质上是偷偷摸摸地恢复了不受限制的概括。因为通过采用恰当的坐标，我们就能描述三维空间的**任何**部分（空中走廊就是一例，它根本没有自然边界），或是时空的**任何**部分，或是流体的**任何**部分（例如，在1962年5月24日19时30分06.8秒，有个一立方厘米的物体，它顶部的东北端比某条河流中的某一地点的表面低5厘米）。具有本体上的**自由**和统一，恐怕是坐标系在（实际）科学中如此有用的原因之一。

10. 为什么需要这样做

我认为构造一个可以被看作是具备常识的程序，显然最后必然要以这种或那种方式包含形式化和像朴素物理学（当然还有朴素心理学，朴素认识论等等）这样的常识性知识。虽然有些人不同意这一点，例如那些认为简单统一的学习过程最终也许表现出智能的人，但是他们的先于理论的假定与大多数AI工作者所作假定差别之大，使我认为把这些研究看成属于基本不同的领域，更为合适。无论怎样，我不打算在这里进一步讨论这个特殊问题。

然而，方法论的分歧表现为实际的差距。AI内部最流行的观点看来是：为了证明人们关于表述的思想是行之有效的，必须构造一个"完善"的程序。那些展示出令人印象深刻的全部行为的工作系统被当作成功的最终判据。可以说，这一要求非常受重视，以致在许多研究部门，如果一个学生没有完成这种令人印象深刻的工作程序，就很难获取博士学位。正如

本文所述，朴素物理学方案有意回避了这种完善程序的编制。我们的目标是构造一个形式化，由它对**启发方式恰当**的可行推理搜索空间作出定义。确切地说，怎样搜索这一空间的问题，解释程序的控制问题，信息恢复和互联的问题——那些或可称为计算问题的问题，以及数据结构的选择问题，怎样实现快速搜索，编程语言的选择——那些或可称为实现问题的问题，都将被有意地忽略掉。

比较而言，只有很少的 AI 工作者采取类似的方法论立场，然而我仍然认为，为了使表述问题取得实质性进步，这样做是至关重要的。麦卡锡（McCarthy 1977）也有某些类似的论点。

这不仅仅是一个变换策略的问题，尽管这个方面也很重要。更为根本的原因是，在构造一个完善的工作程序时，快速成功往往离不开简化和限制，这样，它就不可能去解决基本的表述问题。这表现在两个方面：

第一，为了在编制 AI 行为方式程序时取得成功，人们的确必须非常小心地选择程序的范围。为了便于处理，必须以某种方式对其作出限制，这通常是相当苛刻的。标准的限制形式是限定程序的工作范围，对推理程序而言，就是限制话题；对自然语言程序而言，是把词汇限制在一个微型世界中；对视觉系统而言，是限制可见物范围；等等。由此可知，程序所需的表述方式，在前述意义上，并不是非常**彻底**的。此外，表述方式在使用一些技术时，常常会依赖于这个所限制的工作范围（常常就是这样做的），这些技术对小的“玩具型”世界是行之有效的，但却不能直接应用于更彻底的场合。上面已提到若干例子。

第二,计算上的有效表述方式容易造成较低的稠密性。这有一个重要原因:一个稠密的表述方式,必然定义出一个大型而又急剧扩展的可行推理搜索空间。如果可用于控制推理搜索的仅有的启发式方法,是弱的、一般的(数字式)启发法(如 MICROPLANNER 中的深度优先搜索,在 KRL-0 中的局部过程调用,等等),那么有效的计算行为就不可能通过这种搜索空间来完成。但是这些弱的、一般的方法其实就是我们所仅知的方法,因此,为了达到计算上的有效,就必须具有稀疏的表述方式。

这两方面的压力结合起来,促使构造出有限范围的稀疏表述方式,它们是精心剪裁而成的,适用于为这一程序的行为方式而选定的任务域中期望出现的特定的全套行为。但是正如我已经论证的,作为可看作是恰当捕获常识性知识意义的表述方式,彻底性和稠密性是它的基本特性。

这是方法论的观点,此外还有一种与恰当性密切相关的观点。AI 中有一种流行的看法:只有程序的恰当行为方式才是 AI 理论成功的判据。〔诚如我在别处指出的(Hayes 1978d),正是这个判据将 AI 同"信息加工心理学"区别开来〕。接受这一点,仅仅是在接受表述的恰当性的某种行为主义判据方面迈出的一小步,即在某一行为方式程序中,这个判据保证了行为的恰当。在已知实现者所用技巧的当前状态的情况下,这个判据将在稠密的、彻底的表述方式之外接纳稀疏的、有限的表述方式。如果程序有效,这个论点就行得通,于是它的表述方式必然恰当地捕获拟议中的意义,因为这正是我们所谓"恰当"的含义。

这一立场所存在的问题是没有考虑尺度的影响,至少在

较简单的形式中是这样的。把那些在许多小型子世界中行为方式恰当的程序叠加起来,这种简单过程是不可能产生大型"世界"中行为恰当的程序的,至少目前在日常的常识世界中是这样。对这个世界无法只作那种齐整的分割,因为还需要有各个部分之间的相互作用。也就是说,根据这个判据对只做积木推理来说是恰当的积木世界表述方式,对在液体、绳索、杆系、磨擦、滑轮等的语境中的积木推理来说,就不怎么恰当了。刚刚提到的对有限范围的表述方式剪裁的压力,是对或可称作形式化上行相容性的阻碍。所以即使像我那样把这一判据作为对 AI 理论的最终检验,我仍然认为,太严格、太急切地(比如说,经过三年研究)应用它,是自取失败。**我们决不要企图通过在小范围内的短期出击而得到一个恰当的常识的形式化,无论我们制造出多少这样的东西。**

或许有这样的反对意见:如果情况确实如此,那么一个稠密、彻底的形式化就**不可能**参与有效的 AI 程序。但是,这种说法是错误的。我已经论证过,弱的、一般的控制方法没有处理稠密、彻底的形式化的能力,显然,我们需要更强有力的控制方法。我曾经在别的文章中介绍过一种思想(Hayes 1973)——也可参看科瓦尔斯基(Kowalski 1977)、普拉特(Pratt 1977)、麦克德莫特(McDermott 1976)、戴维斯(Davis 1976)的文章,这一思想指出怎样得到那种所需的能力:不把控制看作定义一个机制的问题,而认为它本身是一个表述问题。我们需要把**如何**进行推理的知识加以形式化,还要把那个使推理成为可能的有关现实世界的知识加以形式化。这种元信息可自行参与推理过程,但是,它又与演绎式解释程序有着不同的和特殊的关系:它对自己的活动作出描述,而不仅仅为了有利。

在我看来，研制出表达这种元知识的形式体系和相伴的解释程序，是摆在AI面前的最重要的任务之一。但是——这正是当前争论的焦点，在它之中所表达的形式体系和元形式化的结构，都取决于在它之中所表达的那个世界的知识的形式体系和形式化的结构。我们不能在真空中发展元形式化（只要它们不是我们已有的那种弱的、一般启发法的形式化）；关于常识的形式化，我们必须首先得出一些与现实一样复杂的、其演绎特性将由元形式化来描述的例子。

11. 为什么这是可行的

反对朴素物理学方案的另一种意见认为，这是一个无法实现的奢望，我们对形式化的了解不足以胜任如此大型的表述任务，要完成这一任务可能要用几世纪的时间，等等。对于这些反对意见，最终只有一个回答：进行尝试才会成功，而我在这里所能做的只是表明这种乐观态度的原因。共有四个：

第一个，是以我最近解决"液体"问题的经验为基础的，我一直认为这是"表述理论"中最困难的问题之一（Hayes 1975）。使我感到惊讶的是，使得这些重要问题得以顺利解决的，是对（由物理边界定义的）空间区域而不是对液体区域进行量化的思想。这里的关键之点是找到**区分**液态物体的正确方式：一个借以表示这类东西的判据。我相信，对区分判据的类似的考虑，也可能在另外一些簇的研究方面取得很大的进展。例如在认识的形式化方面，麦卡锡（私人通信 1977）已开始采用

一个以"概念"区分为基础的新方法，亦即对人们头脑里的思想作出区分。

乐观主义的第二个原因，来自前已略述的关于历程的思想。我认为，由于对变化和动作所取的不恰当的本体论，多年来使得物理世界的形式化遇到阻力，而历程将提供一个绕过这一主要障碍的方法。第三个原因基于已经论述过的不编程的方法论。直截了当地说，几乎没有人**试图**建立一个大型的、启发方式恰当的形式化。我们可能发现，从完成行为方式程序的前提中解放出来之后，它会比我们想象的要容易些。

第四个原因是，有一个显然可实现它的方法论。近年来，在许多领域内，已证明这一方法论是相当成功的。

12. 怎 样 去 做

有一个经过验证的正确方法，可以把知识从人头脑里取出来并变为形式化。在 AI 中，费根鲍姆(Feigenbaum 1977)把它称作"知识工程"，然而，语言学家基本上也在使用同样的方法。其做法如下：对"专家"作咨询时(专家就是那些脑袋里装有知识的人；人们知道这一点，因为他能完成人们所关切的任务)，根据专家对他们头脑里的知识所作的内省说明，建立一个初步的形式化。然后，该形式化以特定的方式**实施行为**，并将它的行为方式同专家的做法加以比较。一般情况下，它的行为方式相当差劲。专家仔细观察了形式化的行为方式之后，常常能够更准确地指出第一次内省说明中的不当之处，并能提出更详细的修正版本。这个版本又经形式化、评判和修

正,如此进行下去。专家不断地面对他所作内省的形式结果,随着时间的推移,一般来说,他变得更加善于作细致的内省。

在"知识工程"中,专家是某个方面的专业人员,而形式化往往是一些条件—动作规则的集合,这些规则可以根据一个合适的解释程序来运行,在某种意义上说,这解释程序是一个很标准的程序。在语言学中,形式化是某种使句子分属各种句法结构的语法,专家则是一个讲母语的人,而且,专家常常就是语言学家本人。在这两个领域中,已证实了这项技术十分成功。

我相信,这个形式化、与直觉作对比和修正的过程,也可用于朴素物理学的发展。从所要求的来看,这是一个我们大家都是专家的领域。而在这里,形式化的**行为方式**就是它支持的推理模式。当"专家们"认为,所有而且只有直接的、可信的结果皆出自于形式化公理的时候,行为方式才是恰当的。(事实上,这是恰当性的弱概念,较强的概念是:从可信的结果**所得的推论**也是可信的。若要使用这个较强的概念,就会引起一些棘手的方法论问题,因为它要求人们具有"二阶"内省。对于语法理论,语言学也有一个正相类似的强恰当性概念,同时也遇到完全相似的方法论困难。)看来,多请几位"专家"才可靠,因为若单独工作,很容易忽略某些显而易见的特征。

取得进展的理想方法是建立一个委员会,每个成员分配一个似乎是簇的东西,他必须设法使它形式化。他们相互讲明需要其他簇的什么东西,比如"历程"簇需要某些"形状"概念,而"组合体"簇需要某些"历程"概念,等等。零碎的形式化频频聚集在小组会中,听取另一些(起常识"专家"作用的)成员的批评,同时还要经受恰当性的检验。可以预料,在这些聚

会中，有些簇将解体，而新的簇又将出现。

形式化一开始所需完成的不过是一些精心措词的英语句子。例如，在不对任何东西做实际的**形式化**的情况下，就能够在本体论问题上取得重要进展。然而，要不了多久，就必须从形式上对直觉作出表达。在这里，我想人们应在允许自由选择形式语言方面宽容一些。不少人认为框架式标记法比较适宜，另一些人则喜欢语义网络，等等。没有理由禁止使用这些表面上不同的一阶逻辑，甚或更加奇特的形式体系。唯一重要的是需要使各种形式体系之间的推理关系变得明确。在实践中，这就意味着它们都应当能够转换成谓词演算，但是这不成问题，它们都能做到。一个更需认真对待的观点是：一些特定的簇可能提出特别的专用表述方式。例如形状可以用图形表述。人们可以构想一个簇，用某种特异的方式来表述，其内部的推理关系从外部是不能掌握的，但是只要定义出从它自身的一部分到参照形式化（一阶逻辑）的转换，比如说对相对位置和方向作出陈述，它就与形式化的其余部分相交接。防止这种事情发生是困难的，人们大概也不打算这样做。但是这里潜伏着重大危险，因为对于整体形式化中可能存在的相互作用方式，这一处理方法过早地作出判断，同时这种方法会把一个严重的错误掩盖得难以发现，甚至更难纠正。

得出概念簇的方式还有若干种。例如：查阅辞典；选择某一个专门领域（如烹调法，各种物质的体积测量），并尝试对其进行描述；详细地分析某一日常行为（例如铺床单时抓住两只角，并轻轻拂动：这样做为什么行？）我希望这些，以及其他方面的事情，可以作为有益的起点。

13. 这是科学吗？

可能有这样的反对意见：试图将知识形式化是件抽象的事情，即脱离了特殊感觉通道或任务域，是非科学的，因为不存在成功或失败的明确判据。对它来说，失败会是什么样的呢？如果这个问题得不到回答，朴素物理学就只不过是纸上谈兵。

我认为这个反对意见提得好，对它的回答应当比我当前所能作的更为恰当。问题是，人们可以不断进入形式化的某些领域，谁会说人们到达的地方还不够远吗？我想人们只能说，人的普通直觉具有指导意义。如果有一些"明显"的物理事实不能从公理中"简单"或"自然"地得出，就尚有更多的工作要做。人们可以运用通常对"精致"、"经济"等所作的科学判断，与作为对手的形式化进行比较。(所有加引号的词都迫切需要进一步的讨论，但我不打算在这里进行。)值得一提的是，语言学正好也处于相同的状况，并且经常在方法论的钩子上痛苦地扭动着：可以肯定，对物理可信性和基本的因果关系的判断与讲母语的人所作的语法判断同样可靠；它们确实在很大程度上独立于文化和语言的边界，所以它们作为原始数据很可能是更加可靠的。(这一点或许提醒我们，应当借助于语言能力/行为方式的区别，使自己免受行为主义的反驳，但是有关这种手法的尝试，确实存在一些深层问题。)

如果朴素物理学能够与皮亚杰心理学关于童年时期物理概念发展的大量现成资料建立更为密切的联系，那就太好了

(虽然这些资料看来没有不存在争议的)。当然,与这些资料的相容性会对朴素物理学的形式化造成限制。但是这是一种很弱的限制,因为这些资料与许多不同的发展理论是相容的,通常它们的相互区分也不是十分详尽的(参阅普拉兹尼对此所作的评论,Prazdny 1978)。我所希望的是,朴素物理学的构造或许揭示出概念框架系中某种发展变化的新机制。目前这种情况已在某种程度上出现,因为建造一个形式化,常常是对已有的局部形式化所作的一种(正是在正确意义上的)发展。[1]

参考书目

Anderson, B., *et al.* (1972). 'Beyond Leibnitz.' *Memo AIM*, Stanford AI Project.

Binford, T. O., *et al.* (1976). 'Computer Integrated Assembly Systems.' *Memo AIM-285*, Stanford AI Project.

Bolles, R. C. (1976). 'Verification Vision within a Programmable Assembly System.' Stanford AI Memo No. 295.

Bundy, A. M. (1978). 'Exploiting the Properties of Functions to Control Search.' (To appear.)

Davis, R. (1976). 'Applications of Meta-level Knowledge to the Construction, Maintenance and Use of Large Knowledge Bases.' *HPP Memo 76-7*, Stanford University.

Feigenbaum, E. A. (1977). 'Themes and Case Studies of Knowledge Engineering.' *Proc. 5th IJCAI Conference*, pp. 1014–29, MIT.

Goodman, N. (1966). *The Structure of Appearance*. New York: Bobbs-Merrill Co.

Hayes, P. J. (1971). 'A Logic of Actions.' *Machine Intelligence 6*. Edinburgh:

---

[1] 这篇文章是在日内瓦语义与认知研究所作休假访问时写成的。十分感谢邀请我前去的所长 M·金夫人,以及参加星期四讨论会的全体成员,特别是 M·金、G·特拉特厄和 H·韦默斯。与 M·辛克莱关于皮亚杰研究的交谈,也使我受益颇多。我妻子杰基打印了手稿的几份草稿,她还是可靠的、健全的、常识性直觉的不竭的来源。

Edinburgh University Press.
—— (1973). 'Computation and Deduction.' *Proc. 2nd MFCS Symposium*, Czechoslovakian Academy of Sciences.
—— (1975). 'Problems and Non-Problems in Representation Theory.' *Proc. 1st AISB Conference*, pp. 63–79, Sussex University.
—— (1977). 'In Defence of Logic.' *Proc. 5th IJCAI Conference*, pp. 559–65, MIT.
—— (1978*a*). 'The Logic of Frames.' [In D. Meitzing (ed.), *Frame Conceptions and Text Understanding*, pp. 46–61. Berlin: Walter de Gruyter, 1979.]
—— (1978*b*). 'Naïve Physics I: Ontology of Liquids.' Working Paper 35, Institute for Semantic and Cognitive Studies, Geneva.
—— (1978*c*). 'Naïve Physics II: Histories.' (In preparation.)
—— (1978*d*). 'On the Difference between Psychology and Artificial Intelligence.' *AISB Bulletin*. (To appear.)
Kowalski, R. A. (1977). 'Algorithm = Logic + Control.' Memorandum, Imperial College, London.
Landin, P. J. (1970). 'A Program-Machine Symmetric Automata Theory.' *Machine Intelligence 5*, pp. 99–119, Edinburgh: Edinburgh University Press.
Luger, G., and Bundy, A. M. (1977). 'Representing Semantic Information in Pulley Problems.' *Proc. 5th IJCAI Conference*, p. 500, MIT.
McCarthy, J. (1977). 'Epistemological Problems of Artificial Intelligence.' *Proc. 5th IJCAI Conference*, pp. 1038–44, MIT.
—— and Hayes, P. J. (1969). 'Some Philosophical Problems from the Standpoint of Artificial Intelligence.' *Machine Intelligence 4*, pp. 463–502. Edinburgh: Edinburgh University Press.
McDermott, D. V. (1976). 'Flexibility and Efficiency in a Computer Program for Designing Circuits.' Ph.D. thesis, MIT AI Lab.
—— (1977). 'Artificial Intelligence and Natural Stupidity.' *SIGART Newsletter*, pp. 4–9. Repr. in J. Haugeland (ed.), *Mind Design*, pp. 143–60. Cambridge, Mass.: MIT Press.
Mackworth, A. K. (1977). 'Consistency in Networks of Relations.' *Artificial Intelligence* 8: 99–118.
Pratt, V. (1977). 'The Competence-Performance Distinction in Programming.' *Proc. 4th ACM Symposium on Principles of Programming Languages*, Los Angeles.
Prazdny, K. (1978). 'Stage Two of the Object Concept Development: A Computational Study.' Memorandum, Essex University.
Schank, R. C. (1975). *Conceptual Information Processing*. Amsterdam: North-Holland.
Ullman, S. (1977). 'The Interpretation of Visual Motion.' Ph.D. thesis, MIT.
Wilks, Y. A. (1975). 'A Preferential, Pattern-Matching Semantics for Natural Language Understanding.' *Artificial Intelligence* 6: 53–74.
—— (1977). 'Good and Bad Arguments about Semantic Primitives.' Memo 42, AI Dept., Edinburgh University.
Winograd, T. (1972). *Understanding Natural Language*. Edinburgh: Edinburgh University Press.
Zeeman, W. P. C. (1962). 'The Topology of the Brain and Visual Perception.' In K. Fort (ed.), *Topology of 3-Manifolds*. Englewood Cliffs, NJ: Prentice-Hall.



纯粹理性批判

D·麦克德莫特*

P·海斯在1978年发表了《朴素物理学宣言》。(此文最终得以刊印在霍布斯和莫尔1985年主编的正式出版物中,Hobbs and Moore 1985)在这篇文章里,他提议为了使常识性知识形式化,可采用一阶逻辑作为标记法,为此需要作一番全面彻底的工作。这种努力可以追溯到更早的研究,特别是J·麦卡锡的工作,不过海斯方案的提法既新颖,又雄心勃勃。他认为,使用塔尔斯基的语义学,我们就可以摆脱计算机程序的限制,去研究内容庞大的知识表述问题。这个提议鼓动起一伙人,去实际地尝试用谓词演算的方式来记录所有(或大部分)常识性知识。海斯以自己关于"液体"的文章(也在霍布斯和莫尔1985年主编的书中)开始了这项工作,试图为现实领域确立本体论和标记法,这是一项很诱人的工作。这以后,有多篇同样思路的文章发表(Allen 1984; Hobbs 1986; Shoham 1985)。我本人也曾是这一运动的热心鼓吹者,写过泛泛的宣传文章(McDermott 1978),也尝试做过一些具体工作(1982,1985)。我甚至以海斯的思想为蓝本与人合著过一本教科书(Charniak and McDermott 1985)。

由是,我是怀着切肤之痛来写这篇文章的,文中对于海斯的计划目前已有的和期待得到的进展,基本上取否定的态度。简言之,我认为目前看到的这种微不足道的进展,决非偶然,事实上,未来也很难有大的改观。其原因是,海斯在论证中没有言明的前提,即大量的推理可以作为演绎的或近似演绎的来分析,是错误的。

我希望不要把我在本文中所说的看作是对 P·海斯个人的批评。原因很简单,他并不是唯一的站在我所批评的那种立场上的人。下文中,我将把这一立场称为"逻辑主义"立场,它确实是数人合作的产物,其中有 J·麦卡锡,R·莫尔,J·艾伦,J·霍布斯,P·海斯,还有我,当然,这些人之中,海斯是最能言善辩的。

1. 逻辑主义观点

我首先来介绍一下逻辑主义的立场。让我们从一个几乎每一个从事 AI 的人都能接受的前提开始:程序必须以大量知识为基础。即使是学习程序,在开始时也必须知道许多东西,比它将来所学的还要多。接下来,要假定这些知识在程序中应当以某种方式来表述。这也是几乎每一从事 AI 的人都

---

* D·麦克德莫特所著"纯粹理性批判"一文原载《计算智能》(1987,3),第151—160页。加拿大国家研究会允许重印。

本文是对 1985 年 11 月在新英格兰 AI 学会会议上发言的一个扩展。文中使用了若干可能招致反对的阳性代词。我提请注意,这些代词用来指未具名的人,并不排除此人是女性的可能。

D·麦克德莫特(Drew McDermott),耶鲁大学计算机科学系教授。

能接受的，但稍微有些保留，下文中还要再讨论这一点。

下一步是作出论证：在编写程序本身之前，我们能够而且应当将程序必备的知识记录下来。我们知道这知识是什么，它是人人都知道的关于物理学、关于时间和空间、关于人类关系和行为的知识。经验表明，如果我们试图先写程序，知识就会受损。为了得到有效的程序，先写程序会对人们实际知道的内容作出过分的简化。反之，如果我们摆脱了这种操之过急的情况，就可以集中精力考虑实际知识的全部复杂性。一旦得到有关常识世界的丰富理论，我们就能尝试把它体现在程序中。这一理论将为写出具体程序提供必不可少的帮助。

接下来讨论的是，对于记录知识来说，数理逻辑是一种理想的标记法。从某种意义上说，这也是唯一的标记法。我们所使用的标记法，对于使用者和阅读者来说必须是可理解的，所以它必然包含语义学，而且必定是塔尔斯基的语义学，因为别无选择。标记法的句法并不重要，所以我们就采用传统的逻辑标记法，因为我们已经知道怎样把它扩展到我们希望达到的任何程度。根据设想，较新的标记法(例如语义网络)的优点是以混淆和幻想为基础的——有关实现方式与内容的混淆，以及有关我们有时希望标记法具有某种意义的幻想。

关于逻辑在最终写出的程序中起什么作用的设想，有可能被误解。因为我们在开始时说过，知识应当表述出来，所以人们或许会得出这样的结论：我们所构造的公理理论最终会明显地出现在程序中，并伴以某种解读公理的解释程序，以决定该做什么。事实上，这种模型可能实现，也可能实现不了。例如，可能有这种情况：一个并未明显地出现在程序中的传递

公理，会体现在某种图形移动器中。逻辑主义的观点是，我们应当暂且(比如说一代人的时间)忽略编程序的繁锁细节，代之以记录人们了解的(或一致地相信的)日常生活。如果我们真的能够提出非形式知识的形式化理论，那么毫不夸张地说，我们就得到了我们期望未来编程者所必备的那种东西。从这一观点来看，为什么逻辑主义者偏爱经典逻辑，而不喜欢较新的标记法，如联想网络，就更加清楚了。人们所说的逻辑与其他标记法之间的差别是，后者通过程序以各种不同的方式对知识进行组织，以供使用，而这正是逻辑主义者不感兴趣的。要知道，他们想要的只是事实，像"液体离开倾斜容器"之类的事实。遗憾的是，这并不是他们偏爱逻辑的唯一原因。在其观点中还隐含着一个前提：有相当大量的思想是演绎的。离开这一前提，如下的看法就缺乏基础：你只管记录人们所知道的内容，而不考虑他们是怎样使用这一知识的。

我们如何被设想知道我们对知识所作的形式化何时取得了进展呢？设想我们写出了一些公理。我们怎么知道什么时候我们写出了人们对主题所了解的大部分内容呢？那就是我们再也想不出还要说什么的时候。但是，人们就一个主题写出他们所知道的每一件事情时，并不是那么完美无缺，他们往往会遗漏一些事情。我们怎么知道什么时候确实达到目标了呢？在我看，逻辑主义者是想当然地认为，当所有的直接推论都来自已经写好的公理时，我们就成功了。海斯在某处说过，如果某件事在人们看来是显然的，它必然有一个简洁的证明方法。要使这一说法有意义，人们所作推理必须大部分是演绎的。水从倾斜的杯子中流出；这个杯子是倾斜的；因此，水就会从杯子中流出。如果大多数推理都符合这种演绎模式，

那么逻辑证明方法就提供了一些理想化的推理模型。并不是我们在思想中必须有一个特定的加工模型，因为我们构想的每一个加工模型，都是这一理想推理方法的近似，因而必然同它所许可的推理方式相一致。我们在设计推理机制时，实际上可以通过所得到的演绎根据来记录这些推理的理由。这一思想潜在地产生出数据依赖性(Doyle 1979)：为了得出结论，程序可以使用任何方法，但是必须能够列出所有证明它为正确的实际前提，所以如果任何一个前提被取消，结论也就取消了。这样，理想化的推理方法就证明了实际推理方法的正确性。

2. 对演绎的辩护

但有很多推理并不是演绎的。如果我发现一个刚刚还盛满汽水的杯子空了，我会推想是我妻子把它喝了，这并不是演绎(福尔摩斯的看法除外)，而是推理的最佳解释。(把推理误认为演绎的唯一方式，是把逻辑编程过程误认为逻辑本身，以后还要再讨论。)如果几乎所有的推理都属于这个或别的非演绎范畴，逻辑主义纲领就陷入困境。情况必然是这样：或者我们想要的推理的主要部分是演绎，或者有多少定理从已知公理集合中演绎而来是根本不重要的。无论从演绎中得到的是什么，在这种情况下，根据事实本身都是微不足道的。

遗憾的是，你越是想推进逻辑主义的方案，你发现演绎就越少。你会发现，情况原来如此：许多推理看起来是那样简单明了，可见它们肯定是演绎，而结果却含有非演绎的成分。我

们可以从海斯(Hayes 1985b)“液体”一文中的两个例子开始：

> 设想一个无渗漏的开口容器是空的，在时刻 t 降水过程开始，降水过程的**底部**就是这个容器开口的顶部，例如，拧开浴缸的龙头，并把塞子塞上。根据公理(46)，这一泄出就会到达另一边，就是浴缸内部向内的那个面。根据公理(59)，浴缸内部必然会产生注水过程，所以根据定理(61)，水的**容量**就会增加。只要龙头一直开着，水就继续增加。我们设想，浴缸最后灌注满了水，亦即达到满载。[所以浴缸将要溢出。](请注意，假如容器是封闭的，比如说是由管子注水的水箱，那么按同一推理思路，必然有一个泄出过程，而这是不会发生的……由这一矛盾可以得出结论……此刻水流的到达必然中止，于是沿着供水管道的流动……必然也停止了……)

这一套论证看起来好像很完善，正是海斯期待的那种圆满解释：清楚的推理具有简短的证明。可惜的是，它们也许是简短的，但却不是证明。设想我们接受海斯对第二种情况的分析，假定注入过程持续了一定时间，出现矛盾，然后得出结论：它毕竟不能持续那么长的时间。然而这样一来，如果我们要遵循某种统一规则，第一种情况就必然是作出假定而不出现矛盾的情况。这不错，但是在演绎中不允许作这种形式的论证：“假定 P；没有矛盾吗？好吧，得出结论 P。”实际情况不是如此。

另一个例子是逻辑主义者处理计划问题的方式，例如罗森斯海因这样的逻辑主义者(Rosenschein 1981)。他们提出如

下问题:如果就现实中的某些行动的结果、事物的初始状态和事物的目标状态给出若干公理,就能找到一个动作序列,并能证明这个序列将现实从初始状态转换到目标状态。这是一个值得研究的问题,但是对于社团负责人、普通人或机器人实际制定计划来说,没有什么关系。想想你最近一次做计划,扪心自问,你是否能做到**证明**这个计划是有效的。这种情况很常见:你可以不费力地举出好几个可信的会使计划无效的条件,但是你还是径直采纳了这个计划。事实上,制定计划的所有难点,特别是在执行过程中对计划的修订的难点,都与把证明计划正确作为目标这种思想不相容。

这一简单的回顾,根据的是包括我在内的逻辑主义者已经得出的一些粗浅的结果。在一个又一个例子中,实际上能够写下来作为公理的东西微乎其微。(这里,我尽量避免引用别人的著作,我自己著作的一个例子是 McDermott 1985。)另一方面,我们看到,像福伯斯一样的非逻辑主义的研究者们致力于编写可得出推理的算法,但是自己却接受逻辑主义者的暗示,认为他们的确应该能够将这些算法中的知识内容表达为公理。其结果显得有些可笑(例如福伯斯文中的公理, Forbus 1984),并没有达到他们预期的那种表达。我常认为这失败属于福伯斯,但是现在我要为他开脱,应责备的是这一任务,看起来行得通,实际上却办不到。[1]

我所说的障碍,对逻辑主义者说来并不是新东西。逻辑主义的方案还有待证明是否正确,或作出专门的解释,这一点

---

① 我说福伯斯受到"暗示",是确有其事的。我查阅过他的文章,并要他尝试采纳更多的逻辑主义。在此致歉。

从一开始就很明确。下面,我将描述所有已知的为逻辑主义所作的辩护,并论证它们都是站不住脚的。这些辩护并不是相互排斥的,每一辩护都对一组别的缺陷作出补充,同时大多数逻辑主义者很可能在大多数时间里对它们之中的大多数信以为真。现列举如下:

1. “理想化”辩护:把问题的演绎公式看作理想化。例如,演绎计划问题可以看作“真实”计划问题的理想形式,并可能在作为现实世界理论的理想化形式的一个世界理论中继续进行下去。

2. “词汇”辩护:强调我们可以选用我们想用的任何谓词。我们无法由演绎推出某一特定计划是有效的,这或许是事实,但是,如果我们将问题转变为推断“应当做(执行者,计划)”的问题,就比较容易取得进展。

3. “科学女王”辩护:找出演绎和非演绎推理之间丰富的结合点。例如,对最佳解释所作的推理,可以看作找出前提,以前提为据,经演绎得出一个明确结论。这样,演绎就变成一个大推理理论的中心,为许多有价值的演绎变异形式所簇拥着。

4. “元理论”辩护:断言存在着演绎的“元理论”。这些理论涉及怎样发现和校订那些原始的、有缺陷的“客体层次”理论的结论。

5. “演绎技术”辩护:以存在着逻辑编程为根据,论证许多现实的推理问题可看作在本质上是演绎的。

6. “非单调”辩护:论证通过扩展经典逻辑,接纳原先不可行的结论,就能获得远比以前大得多的推理集合。

我将逐条对每一辩护作出反驳,从理想化辩护开始。

理想化并非总是不好的，它们常常是很重要的。例如，证明某个计划将在下棋比赛中获胜，很可能就是一个有用的理想化，尽管这个证明忽略了可能有人会突然给制订计划者100万美元，让他故意输掉比赛。人们完全有理由把那种可能性根本排除在公理之外。然而我所反对的心态是，假定在所有棋局中，目标都是要找到能证明获胜的策略，或偏爱那些使这一证明成为可能的棋局，而忽略更多的常规棋局。事实上，现实的下棋程序（以及人类棋手）并不去做那种间接类似于证明计划将行之有效的事情。当然，对于任何已知的算法，比如说对策树搜索，从某个角度，也可以认为它是在演绎某种东西（例如树的极小极大值），但是这种观点与我们无涉。

我担心，在很多情况下，演绎问题被说成是对现实问题的近似，而实际上它是对现实问题的对等物，是真实事物最好的演绎模仿。在很多情况下，这个真实不会是不可逾越的。试图写出对等物领域中的事实，有可能带来对实际领域的深刻认识。所得出的本体论和公理，对于最终的程序编写或许是有益的。但是，我们不能指望从理想化中得到逻辑主义者所期望的覆盖范围。来自真实领域的许多概念恰恰不能在理想化中找到。相反，还有一种危险：有过多的来自理想领域的概念无法在真实领域中找到，并且理想化被扭曲到无法使用的程度。但是作为策略，理想化的使用看来还是有价值的，本文末尾还要再谈到这个观点。

下一个是"词汇"辩护。这里所述的论点当然是我所愿意接受的。如果有人正在设计一个能作数学思考的程序，采用演绎方法并不需要把程序词汇限制在策梅罗—弗朗克尔的集合论之中，而是想要得到人类数学家可能使用的那些谓词，像

有价值的概念(C),看来可以证明(定理),如此等等。例如,人们可能选取莱纳在 AM 中(Lenat 1982)隐含地使用过的所有谓词,并尝试编写一个程序,从中演绎得出一个概念是有价值的,或者一个定理大概是可以证明的。

这个辩护的问题是,它始终没有起过什么作用。我们拓宽可作为演绎的问题范围,在许多情况下,只不过是在我们无法用演绎解决的清单中增加了一些问题而已。我们完全有理由说,AM 不是一个演绎程序。

"词汇"辩护的另一个问题是,它允许我们用普通问题代替难点问题。从实例看,在医疗诊断中,如果在演绎"诊断(病人,疾病)"时遇到了麻烦,切换到可能的"诊断(病人,疾病)",是不起作用的。在这种情况下,新问题过分简单,所有行为都存在于鉴别诊断和有分量的证据中,现在将被忽略,或是用某个非演绎的模型来搪塞。

"科学女王"辩护可以这样来详述:考虑"不明推论式"[①],这是 C·S·皮尔士用于解释性假设生成的术语。这个过程是非演绎的,但是我们可以把它看作一种"逆向演绎"。例如,为了解释 q,寻找一个已经知道的"(如果 pq)"形式的关系,并假定 p 是一种解释。更一般地说,为了解释 q,找出与已知事物相结合后必然得出 q 的那些前提。如果这个模型是正确的,那么即使不明推论式不属于演绎,它也仍然可以通过演绎而得到确证。查尼亚克和麦克德莫特对这一观点有保留地表示赞同(Charniak and McDermott 1985)。

这种解释在哲学家们中间被称为**演绎推理法则理论**。它

① 不明推论式是指逻辑三段式中小前提无证明的推论。——译者

总是与C·G·亨普尔的名字连在一起(Hempel and Oppenheim 1948; Hempel 1965)。可惜,别的人几乎都不相信它。它是被作为一种科学解释模型而设计的,但是它有若干缺陷,看来很难作为解释个人或物理系统的行为的模型。这里的问题是,在**解释要素**与**待解释事物**之间的演绎链既非必要,也非充分。

它之所以不是必要的,一个原因是解释只要使所观察的事实有或然性,我们就可以满足。对此,亨普尔是允许的,所有的诊断专家系统,如"麦森"(Shortliffe 1976)和"探测者"(Duda et al. 1980)也是允许的。

但是有很多超出常规的例子。设想你在报纸上看到塞考克斯的塞尔玛·麦克吉利科蒂刚刚赢得了新泽西州的彩票,这是近两月中第二次,每次都超过10万元,并无舞弊之嫌,因此你得出这样的解释:**这是公正的抽彩,碰巧有人接连赢两次**。这是个令人满意的解释,但是不能根据它推出"塞尔玛·麦克吉利科蒂在两个月内赢得两次彩票。"〔W·萨蒙首先指出这一解释类别(Salmon 1967, 1975)。〕

资料的演绎所以不是充分的,原因在于这一条件太容易满足。一般说来,推出已知结论的演绎方式有成千上万,但是作为解释,几乎没有一个是合理的。例如,有一天我发现我的收音机闹钟快了两分钟。因为我特别关心钟表的准确,所以很烦恼,便寻求解释。一个情况是,两小时前电池用完了,因此可以推想钟会慢两小时,但是我记得这是一个有备用电池的钟。所以合理的解释就是带电池电源的钟不准确,每小时约快一分钟。

我们假定这个解释与结论一起可以转化成一个演绎论据。"钟快了两分钟",那又怎么样? 对这同一结论,存在着大

量别的演绎方法。(“一个来我家的访问者恶作剧地把钟拨快了”,“宇宙射线的爆发正巧击中我的钟”。)人们可以反驳,这些解释显然是牵强的,根据“科学女王”的思想,我们只能寄希望于描述一下恰当解释的特征,但是这是一种没有意义的说法。恰当性的条件恰恰是太平凡了。

如果我们不留心,它甚至会变得更加平凡。只要这只钟是房间里唯一的钟,像“房间里的每一只钟都快两分钟”这样的前提,就能解释“这只钟快两分钟”。亨普尔试图通过规定前提和结论是“定律式的”,来回避这个问题。这一特性等价于什么,完全不清楚,但是,它可用于排除用“行星个数是一个素数的最小奇数平方数”来解释“有九个行星”。实际上,它从思考中排除了对特定事实的任何解释,而把这种理论变成为解释定律的理论。

“女王”辩护在一定程度上是“理想化”辩护的翻版,所以有着类似的弱点。在假设与需要解释的证据之间,必定有某种联系,但是把这种联系说成是演绎的,只不过是武断的臆测。一般地讲,我们就此所知的只能是,一个好的假设就是能让典型人类提问者满意的假设。稍后还要再次谈及这个问题。现在可以得出结论:演绎不能成为不明推论理论的核心。

“元理论”辩护认为,带有演绎推理方法的问题,可以通过引入演绎的“元方法”而确定下来,这个元方法或干预并校订它的输出;或改变它的前提;或使它变得能容纳对立理论。例如,在海斯的两个容器例子中,我们可以设想这个元方法能作出信念修正,引入流动持续性前提,并在棘手的问题出现时,将它们撤回。

这一辩护的问题是内容空洞。推测起来,演绎元理论的

题材必然是“对客观理论的合法干预”。但是在这种理论中，并不存在来自人类直觉或其他什么方面的约束力。可以肯定，根本没有什么约束能使干预保持演绎的完好性。假如有的话，这一辩护就不能实现对演绎的必要强化。所以要弄清怎样排除这样的理论：“在周末相信所有奇数符号的陈述，在平日相信所有偶数符号的陈述”，是很困难的。如果这项计划变成精心制作有这种随意功能的元理论的计划，我们不妨说，我们终究还是在编程序。

关于元理论思想，**在总的方面**，没有什么可说的；然而对于具体的例子，可说的又太多。让我们看看**信念修正**的思想，两段前曾谈及此。如果你开始认真地研究它，最后你就会以研究非单调逻辑而告终（更多的讨论见后）。这一研究会使元理论体系相形见绌。为了取得进展，你不得不构造十分复杂和详尽的模型，而且这与它的目标是完成某种演绎元理论，还是编制 LISP 程序，都全然无关，这一点远在你做完这事之前就很清楚。如果你选择的不是 PROLOG 而是 LISP，那么元理论体系就毫无用处。

这样就到了第五个辩护——“演绎技术”。对演绎的作用很容易估计过高，原因之一是由于它有一组得力的工具，诸如反向链接和统一化，它们来源于定理自动证明研究，但是已在 PROLOG（Clocksin and Mellish 1981）和 MRS（Genesereth 1983）这样的系统中得到较广泛的普及。结果这些工具推出一个精巧的计算模型，它和传统模型一样强有力，在某些情况下甚至更加理想。由于可以用这些模型做任何计算，也由于它们是在定理证明机中产生的，自然会得出这个结论：在一定意义上，任何计算都是演绎。我们很难反驳导致这一结论的论点，因

为这样的论点根本不存在，它只不过是概念之间的模糊联想而已。（谬误当然不会因逻辑编程群体使用“每秒进行的逻辑推理次数”这样的短语指示像表处理操作那样的小事而消失。）严肃的研究者无意中被这个谬误所迷惑，然而他们也可能受到演绎技术精巧性的吸引。

我们以上面批评过的那种含混不清的思维为例，看看在PROLOG类型系统中计算数值的方法。这里，含有变量的目标被转换成找出变量值的要求。目标“系属(〔a,b〕,〔c,d〕,X)”意味着“找到一个X，它是系属〔a,b〕和〔c,d〕的结果”。如果公理写得正确，就能找到数值。在这个例子中，X被限定为〔a,b,c,d〕。可将这一目标同系属(〔a,b〕,〔c,d〕,〔a,b,c,d〕)作对比，后者的目标是要确证〔a,b,c,d〕就是所要的结果。合理地正确运行的PROLOG程序的特性是，只要程序能找到一个数值，就能对其加以确证。〔要得到相反的性质就困难得多了(Shoham and McDermott 1984)。〕

这里的问题是，虽然证实结论的思想一般可以归结为演绎(因为它恰恰就是证明某事物的思想)，但是计算数值的思想却并非如此。从逻辑的观点看，系属([a,b],[c,d],X)只不过是(非(存在(X),系属([a,b],[c,d],X)))的司寇伦式的翻版。(关于司寇伦形式和“非”的解释可参阅任何教科书。)通过找寻X的值，用反向链接证实这种结论，实质上是一种有用的例外情况。如果将其扩展到反向链接之外，这种思想就会完全失败。勒克姆和尼尔森(Luckham and Nilsson 1971)给出一个可用于任何分解证明的变体形式，但是对于每个变量，并非每个分解都生成单值。更重要的是，一旦将逻辑扩展到有限可公理化的一阶理论之外(这种情况常常出现在知识表

述事务之中)，整个分解思想和司寇伦形式就变得不起作用了。

即使在这一思想有效时，逻辑也无法提供一个答案构造的一般性理论。我们来看 R·莫尔的“马桶炸弹”问题：你从邮件中收到两个无法区别的滴嗒作响的东西，另有匿名电话警告说，其中肯定有一个是炸弹。从看过的电影中得知，把炸弹放入抽水马桶是一种可靠的防止爆炸的办法。你该怎么办？答案是，把两个东西都放入马桶。(也许放入浴缸更好。)但是如果我们把这个问题变为如下逻辑形式：

结果(计划，与(消除爆炸(物体 1)，消除爆炸(物体 2)))

我们也可回到(采用勒克姆和尼尔森的做法)：

计划 = 放置(物体 1，马桶)

或

计划 = 放置(物体 2，马桶)

也就是说，在没有实际构造出计划的情况下，定理证明机已经得意地证实存在着一个可操作的计划。当然，我们实在不能再作更多别的要求了，演绎恰恰没有提供计算任意事物的理论，它追求的只不过是证实任意事物的理论。

3. 非单调辩护

最后，轮到最有说服力的辩护——求助于“非单调逻辑”。该名称用于这样一个逻辑系统：有更多前提时，其中一些结论可以废除，亦即可以撤消。这类逻辑看来是为像上面提

到的那两个水箱例子而特制的。我们希望得到的推论是，只要没有理由认为有其他情况出现，水就仍然流进水箱。当矛盾出现时，结论就被撤消。

根据定义，非单调性几乎是与演绎不相容的。因而正如伊思雷尔（Israel 1980）所指出的，“非单调逻辑”的说法有些自相矛盾，很像是为了补偿素数的某些缺陷，而提出研究“复合素数”。在实践中，“非单调逻辑”指的是一个推理系统，它向获得明显的可废除推理的普通逻辑提供了一个简单的一般性扩展。我们并不指望这样的系统可以从推理中得出最佳解释，但是我们完全可以期待它推导出：你的汽车还在你原来停放它的地方。

既然对普通逻辑来说可能存在着许多可选择的“简单、普遍”的扩展，我们就很难对非单调逻辑的发展下任何最后的结论。然而，我们能够考察已经完成的情况，并估计它未来的前途。已经采用的方法主要有两个：**缺省法**和**划界法**。缺省法是将 PLANNER（Hewitt 1969）和 PROLOG（Clocksin and Mellish 1981）的“以否定为失败”的思想形式化。通过引入推理规则：“根据前提 p 和不能推断出 q，推断出 r”，对普通逻辑加以扩展。其思想是，在缺乏特定的压倒一切的信息 q 的情况下，r 是根据 p 得到的有缺省的结论。我们看一个例子

$$\frac{(\text{鸟 a})\text{相容}(\text{非}(\text{反常 a}))}{(\text{a 能飞})}$$

这里，“相容算式”表示该算式与系统中所有的推理都是相容的。这样，对于任何给定的鸟，我们就可以按正常情况推断出它能飞，但是如果存在着用于推导出反常情况的公理，就可以

以它们为"门",把这一规则拒之门外。与这种形式大体相似的系统已由赖特(Reiter 1980),麦克德莫特和多伊尔(McDermott and Doyle 1980),克拉克(Clark 1978),以及其他人作过研究。

划界法是由麦卡锡(McCarthy 1980)和他的同事(Lifschitz 1985; Lifschitz 未刊稿①)发明的,它避免了增加新的推理规则,而主张采用一个带有公理的一阶理论,以表达使某一谓词"极小化"的目标。举例说,现有公理

(对所有(x)(如果(与(鸟 x)(非(反常 x)))(x 能飞)))

若能使反常谓词极小化,就可以像以前那样,对每一个给定的鸟,按正常方式推断出它是能飞的。为了实现这一点,就要在原有理论中增加一个二阶公理。设 A(反常;鸟)是已有的全部公理的合取。(最好是有限的。)我们在 A 后的括号里写出想要替换的谓词的名称。分号把准备被极小化的谓词(反常)与"变量"谓词(鸟)分开;通常可能有一个或多个谓词要被极小化,有零个或多个变量。所以 A(foo;baz)可作为同一公理集,其中 foo 代替反常,baz 代替鸟。根据这种标记法,新的公理就是

(对所有(pb)
　　(如果(与)(A(p;b))
　　　　　　(对所有(x)(如果(px)(反常 x)))
　　(对所有(x)(如果(反常 x)(px)))))

亦即如果(在 A 由于鸟的变化而弱化之后)p 是任何满足 A 的

---

① V·利夫席茨(Lifschitz 1986),"点式划界法",未出版手稿,1986 年 1 月 16 日。

谓词,并与反常具有相同强度,那么反常也与 p 具有相同强度。现在还需另一步骤,就是插入值以代替 p 和 b。假定已知的鸟中只有 Clyde 不能飞,那么 A 就包含着正常鸟会飞的公理,以及(鸟 Clyde)与(非(Clyde 能飞))。如果我们插入 p = (∧(y) (= y Clyde) 和 b = (∧(y) (= y Clyde)),那么,因为 A(p;b)成为

(与(对所有(x) (如果(与(= x Clyde) (非(= x Clyde)))
(x 能飞)))
(= Clyde Clyde)
(非(Clyde 能飞))

我们就得到了划界公理的例子:

(如果(与(对所有(x) (如果(与(= x Clyde) (非(= x Clyde)))
(x 能飞)))
(= Clyde Clyde)
(非(Clyde 能飞)
(对所有(x) (如果(= x Clyde) (反常 x))))
(对所有(x) (如果(反常 x) (= x Clyde))))

但是这关系的前项来自 A(反常;鸟),所以我们可以推出后项:Clyde 是仅有的反常对象。因此,任何别的鸟(如果能表明它不等同于 Clyde),都被判定为是能飞的。

请注意划界法是怎样获得非单调性的。当 A 增加一个新的公理时,划界公理就会变化,同时往往有某个定理不再适用。

对所有各种已知的非单调逻辑来说,存在两个问题。第一,如果不经过很大的努力,一组规则的结果如何,往往弄不

清楚。第二,想要作适当“扩大”,往往会失败;就是说规则的结果过弱。下面我用短语“你不可能认识”和“你不想知道”来标明这两个问题。

在缺省逻辑中,会出现“你不可能认识”的问题,因为一个公式是否与一个理论相容一般是不可判定的。事实上,当所讨论的理论是一个包含缺省规则的理论时,甚至对短语“与理论相容”的含义作出准确定义都是很困难的。在推导出每件事情之前,你不可能辨别什么是无法推导的,所以我们转向了缺省理论的“稳定扩展”或“固定点”的思想。这样的固定点是一组公式,它在直觉上是“稳定信念集”,其特征由“非结论公式集”来确定,这样,(a)固定点上的每件事,经由缺省推导规则,都来自原有理论及非结论公式;(b)没有非结论公式随之而生。如果已知 Clyde 是鸟,那么(非(反常 Clyde))就是一个非结论公式,因此(Clyde 能飞)就是固定点的一个元素。除非(反常 Clyde)也是可演绎的,这时(Clyde 能飞)不在固定点上。遗憾的是,这些固定点和非结论公式集是无穷的,一般情况下,很难描述或找全它们。

划界法使用起来也有困难。所有各种已知的划界法共同具有的是达到结论的这一步骤:

·在原有理论上增加二阶公理

·猜测谓词常项,插入到公理中去

·化简

这正是我举例时采用的那种步骤。它的问题是,新增信息的数量往往与最终想要得出的结论相同。事实上,它考察的东西往往与那些结论相同,另外带有几个附加的 λ。

从原理上讲,划界法可以机械地使用,转动一下曲柄,所

有的结论都出现了。但是在实践中,没有一个方法能够枚举出二阶公理的有用实例,所以划界法只用于一些不起眼的问题,这些问题要得出的结论是原本知道的。(在特殊情况下,可以证明,划界法和缺省逻辑都可以还原成可计算的算法,但是我们这里对这些特殊情况不感兴趣。)令人费解的是,无法判定的缺省逻辑提出了一个有某种实际效果的具体算法:为了确证 p 是相容的,尝试证明它的否定形式,并失败。在这一步骤中止时,它往往成为富有启发性的近似方法(Clark 1978)。

非单调逻辑的难题导致了一个奇怪的现象。逻辑主义者大步向前,使用非单调构造,并在有关文章中宣称,他们所希望的哪些结论会出现,而实际上他们不知道这些结论是否会出现。在这一点上,他们所描述的推理在何种意义上可由形式系统加以证明,已不复清楚。既然希望成真,带有希望的思考也就无关紧要了。遗憾的是,它把我们带向“你不想知道”的问题:认真研究一个非单调系统时,常常有这种情况:形式系统实际允许的结论与所预期的不同,通常是更弱一些。在缺省公式中,由于上述固定点常常是非唯一的,该问题就会出现。其中有一些是合理的,但是还有许多对应于信念集,是制定基本规则的人所排斥的。(这些固定点无法通过增加更多的缺省规则来消除,那样只能使事情更糟。)如果一个理论具有数个可选择的固定点,实际上什么能被说成是这个理论所含的定理呢? 或者,这样的理论不含定理,它们无法作为我们正在寻找的理想化的推理工具;或者,我们坚守定理的弱概念,即定理是在所有固定点上都可推导出的某种东西。常见的情况是,这种供选择的点给出一些选言定理,其中某些选言

命题是来自多余固定点的、反直觉的冒牌货。我们希望得到结论 p,但是我们最终得到的是 p **或** q,而 q 超出了正常范围。

这种过弱的选言命题,也会不期地以划界形式出现。对于划界法来说,这一现象有些不同,因为固定点的概念并不具有同样的证明论作用。不过,我们有一个与之对等的思想:**极小模型**,其定义如下:就某个谓词 P 而言,如果一个模型与所有其他(非变项)谓词无矛盾,且它的 P 是另一模型的 P 的子集,它就比另一个模型"小"。极小模型是指没有更小模型的模型。可以证明:一个公式在 A(P;V)的所有极小模型中为真,如果它由 A(P;V)和上述二阶划界公理得出的话。

现在,过弱的选言命题问题出现在下述形式中。常见的情况是,会出现一些在重要方面有差别的极小模型,以致有些模型对人类观察者来说是"显然错误的"。在句法方面,划界法会产生选言命题,使每一选言命题刻划一类极小模型的特性。因此这一状况与缺省逻辑实无差别,只是选言命题是作为基本机制的结果而出现的,并不是作为对定理概念的一种拼凑起来的定义方式硬塞进来的。

汉克斯和麦克德莫特在最近一篇文章中(Hanks and McDermott 1985,1986)对这一现象的实例作了详细的考察。我们研究了麦克德莫特有关时态逻辑的简化形式(McDermott 1982),这一形式比先前研究过的非单调系统更加复杂些。我们曾希望证明,从形式系统中真的能够得出我们想要的结论。我们曾期望多重固定点问题能胜过缺省逻辑,但是我们也曾期望划界法是有效的。我们惊异地发现,划界法和缺省公式面临同样的问题,其实回想一下,各种系统之间的相似性是非常显著的,所以不必对此感到奇怪。

所有的逻辑都有这个问题:像“极小化”和“稳定信念集”这样的概念恰恰不适合于时间域。我们想要的非单调规则曾是(非形式地说)“世界的状态趋于保持不受扰动”。所有逻辑得出的结论都使扰动达到极小,但这并不是我们真正想要的。事实上,我们希望的是避免未知原因带来的扰动。

我们曾试图说明:“历程”如果不被后续事件明确“截断”,就会继续下去。看看下面的事件序列:

1. 弗里德出生了	弗里德开始成为**活着的**
2. 枪装上弹药	枪开始成为**装有弹药的**
3. 弗里德被击中	弗里德成为**死的**

我们应该能够得出结论:弗里德现在死了(不幸由于暴力)。但是另一情节能同样成功地使干扰达到极小。在这情形中,事件3发生之前,枪停止装弹药,而这只是为了避免对弗里德的存活造成干扰,别无他因。

那篇文章之后,麦卡锡小组成员V·利夫席茨已经证明[①],“点式划界法”这一新思想将解决汉克斯—麦克德莫特问题的简化形式。无人知道它是否能解决更复杂的形式,更谈不上解决实际的物理学公理集。无人知道还会出现什么其他问题。但是这个“解”的真正麻烦,是它比起以前各种划界法来,甚至更加头重脚轻。我们必须知道答案,只有这样,划界法才能为我们确证答案。此外,允许谓词以“极小化”类存在于它们的部分区域中,而以“变量”类存在于另一些部分中,因此必须用附加关系的形式提供信息,说明哪一部分到底是什么。

---

① 见本书第295页注①。

这个解法为保全划界法而毁坏了它。正如所有的划界法形式一样，我们从希望用来扩大我们的演绎理论的结论开始，而找到的却是一个为我们提供这些结论的二阶公理。如果我们选出的第一个公理无效，我们就会发现另一个公理。一旦任务完成，我们就丢弃这个公理。我们全然不知怎样在我们曾论证的结论之外引申出任何别的结论。在这种情况下，公理起到什么作用呢？怎样才能说，是它在证明所期望的结论，而不是所期望的结论在证明它呢？在实践中，直接在理论中增加那些结论，并非难事。这一过程有可能每一步都是非单调的（若理论变化，改变的只是新增的内容），同时每一步都变幻无穷。

我们原来的目标是简单、普遍地扩展经典逻辑，使它产生出"显然正确"的结论，然而这一目标却难以达到。在缺省公式的情形中，原因在于这种诱惑产生了非递归可数定理。在划界法的情形中，原因在于我们必须在得出答案之前提供答案。在这两种情形中，即使能得到答案，也常常太弱，虽然在划界法中，我们常常有权切换到不同的划界公理。

这一危机对于作非单调推理的程序并未产生什么影响，认识到这一点很重要。所有计算机化的推理几乎都是非单调的，因而也是非演绎的。这是我们一开始就遇到的问题。这个危机所影响的，是我们尝试把演绎稍作扩展以包含一些"显而易见"的情况的工作。现在的状况是，并不存在一个非单调系统，可以证明我们的程序所作的非单调推理是正确的。相反，最后的结果是，我们不得不花费大量精力来改造形式系统，以重复简单的推理过程。这种努力对于编制程序来说，只是个附带问题或事后之见，没有作出什么实质性贡献。

如上所述，情况也可能改善。今后有人可能发现我们正在寻找的那种非单调系统。但是从目前看，我们只能得出这样的结论：不能把非单调性看作解决某种演绎问题的方法。

4. 没有演绎也行

让我们对前面的论述作一总结。我举出了逻辑主义者的方案：以逻辑公理形式来表达常识性知识。我概述了为这一方案所作的辩护，并指出其隐含的前提：大量的推理都是演绎的。我论证了这一前提的错误，即使用各种方式对逻辑加以扩展，也是无用的。

离开这一前提之后，原来的论点会出现什么情况呢？不难看到，无论就一个领域写出多少公理，所期望的大多数推论仍然不能从它们之中推出。为了推出这些推论，还必须提供一个**程序**。换言之，在大多数情况下，不可能离开"加工模型"去建立一个"内容理论"。（这些术语出自L·伯恩鲍姆。）一般认为，内容理论是关于人们知道什么，他们怎样划分世界，他们的"本体论"是什么的理论。而加工模型则解释人们是怎样使用这个知识的。内容理论处在纽厄尔（Newell 1981）的"知识层次"上，恐怕与怎样处理它所表达的事实无关。我们现在可以得出结论：如果所考虑的推论是非演绎的，那么内容理论的用途是有限的。你不能只是开始罗列人们知道的事实，把它们用逻辑或任何别的标记法表达出来，而不说一下你如何假定它们将被一个程序采用，因而要尝试说明何种推论。你可以避免这种麻烦的唯一情况是，你能够指出一类有价值的、

包含这一知识的纯演绎推理。在这情况下,你对于每一候选的加工模型的确有了充分的了解,不再需要说什么了。但是从目前看来,这种推理是极为少见的。

顺便提及,正像对逻辑主义方案一样,这一点对莱纳等人(Lenat et al. 1986)的 CYC 方案也完全适用。他的小组利用了范围更宽的工具,彻底放弃了逻辑的学科准则,但是同样的异议又出现了:他们怎样知道什么时候才算是取得了进展呢?

反对独立内容理论的论点,给原来赞成塔尔斯基语义学的论点带来不利的冲击。若没有程序,指称语义学就是规定我们的标记法的意义的唯一方法。但是关于知识表述,存在着与它对立的传统思想,认为知识表述系统在本质上是一种有特定目标的高级编程语言。这一观点在如 OPS5(Brownston et al. 1985)和 PROLOG(Clocksin and Mellish 1981)这样的系统中描述得很清楚。但是它也用于许多联想网络,它们常常用作为 LISP 码组块的工具。其实,OPS5 和 PROLOG 还不是主要的例子,因为它们是通用的编程语言。更好的例子要算语法分析规则标记法,如马库斯(Marcus 1980)采用的那种。虽然这个标记法表达出关于语言句法的"知识",但是它不含有指称语义学。它的语义学是**过程的**;如果一组规则能使语法分析得出正确结果,它就是正确的。

换句话说,相对立的过程传统是,知识表述系统实际上不表述任何东西。这种立场使得典型的逻辑主义者大为震惊,因为这就意味着承认被表述的知识本质上只能以一种方式被利用。要数出各"方式"是件难事,但是可以描绘两种不同模型所需要的"同一事实",其中每一个都带有它自己的专用编程标记法。这事实必须表述两次。这当然是对于智能程序的

一个令人不快的要求。

若是标记法同时具有指称语义学和过程语义学，那就好办了。这并非不可能，任何被程序实际使用的以逻辑为基础的标记法，**根据事实本身**就是具备这种双重语义学的。（纯PROLOG程序即为一例。）有人想作相反的推测：任何过程标记法都可以转换成等价的指称标记法。这岂不就是清除一些不相容性并形成某种本体论吗？可惜，这种乐观的估计是建立在对实际上怎样运用如联想网络这样的工具的错误观念的基础上的。在一些研究者心里，认为标记法具有某种形式语义学，并且不大怀疑存在着看起来更像传统逻辑的等价标记法。但是在大多数使用者心里，系统是某些特征——恶魔等等——的集合，如同标准的编程环境那样（只是他们希望更奇异些）。任何运用这些特征的方式，只要是实现直接编程目标的，都是合法的。关于这点，没有可疑之处。对于系统被滥用的每一研究者来说，很多人会**鼓励**这种对他们系统的"创造性"使用。任何这样一种系统能够找到指称语义学的机会，可以说是微乎其微。

当然，这一令人不无遗憾的实际标准本身，还不足以阻碍我们寻找同时含有指称语义学和过程语义学的标记法。问题是，从目前来看，这一追求没有任何理论基础。有些人坚持认为他们的标记法含有指称语义学；另一些人（人数甚至更多）不愿接受这一限制。不管我这里的论点是什么，我认为自己在性格上仍属于前一组。如果一个学生带着无指称的表述来见我，真让我伤脑筋。以前我认为有理由说服他重作思考，但现在我只能处于理解不良的位置。学生总是能指着他的程序说，这程序并没有从荒谬的标记法中得出荒谬的结论。我要

是得出荒谬结论的话,问题出在我身上。

举一个我爱用的例子。我们来看这样一个简单的事实:“俄国人已经把战舰部署在靠近美国海岸的地方”。除非我愿意采纳如“当前—部署—靠近—美国—海岸”这样一些“德国计算机”式的谓词,否则要对这类内容作出恰当表述,就必须明确地表示出什么是美国海岸,那里大约有多少船只,已发现它们分布如何,意指多长时间,等等。然而,谁来确定这是“恰当的”?用德国计算机,任何特定的应用程序大概都能过得去。许多雄心勃勃的标记设计者会借助于“英国计算机”,如

(俄国人已经(部署军舰(靠近(美国海岸))))

在我看来,这样更糟。但是为什么呢?如果程序有效,毛病在哪里呢?

因此,在原有的逻辑主义论点中,第二步上存在着漏洞,即主张知识必须表述出来。虽然多数从事 AI 的人会同意这个主张,但是我们现在看到他们大多没有这个意思。他们的基本想法是:我必须写出许多存放知识的程序,同时还需要专门的高级标记法来完成它。

这一事实可使逻辑主义者感到安慰:他们的对手要煞费苦心地把那些构成表述方式的“高级”程序与任何旧的程序区分开来。如果无法作出区分,那么所有程序都可以用来“表述知识”。在旧的过程说明争论中,我把这一立场称做过程主义立场。这一争论沉寂下来,是因为没有人对“表述”的这一意义真正感兴趣。例如,按照这一意义,可以说视觉程序是表述有关图像形成的物理学知识的。看来还存在一个更强的意义,在这一意义上,AI 程序是对物体和事实的明确表述进行操作的。指称语义学对这一意义是什么提供了答

案，但是我们现在看到，这一答案对许多 AI 研究者来说毫无吸引力。

5. 对过程的辩护

逻辑主义者并非从不打算编写程序。只是他们期望在程序被写出时，这些程序将被看作定理证明者中的最佳文本。为了证明这些程序的正确性，只需表明这些程序是忠实于作为程序基础的公理的。

这套设想现在既已被否定，我们就需要有确证推理程序的新方法。AI 程序由于存在着难以逾越的复杂性而名声不佳。有时这个特点又被当作优点，似乎智能的神秘性应当保留在它的计算模型中。可是，一个我们无法理解的模型，根本就不成其为模型，尤其是当它只能对个别例子成立时（Marr 1977；Birnbaum 1986）。

要使“确证”的思想精确到足以支持那种认为每个程序都必须得到确证的意见，大概是不可能的。然而，在程序之外，我们能够指出一个明晰、独立的理论，说明这程序为什么有效，这总是令人满意的。例如，在视觉研究中，关键的一步正是从带有神秘气氛的“异性基”模型转变到由物理学和心理物理学所确证的模型。在想象性质的范围内（de Kleer and Brown 1985；Forbus 1984），已经写出的程序并没有什么错处，但是经过凯珀斯（Kuipers 1985）对它们的意义和限制进行了分析，就更加清楚了。

但是，有许多大类的程序缺乏任何理论依据，特别是那

些与推理的最佳解释或**不明推论式**有关的程序。假如我们能回到哲学家那儿，将他们的智慧重新发掘，那就好了。既然他们能创造出像演绎推理那样伟大的理论，毫无疑问，在其他理论上，他们肯定也会做得同样好。令人遗憾的是，哲学家们使我们失望。不明推论式理论或许可以从回答这些问题入手：

哪些种类的事物需要解释？

什么可以起到解释的作用？

什么可以作为解释的证据？

如何衡量证据的支持强度？

证据何时能强到足以证实对假设的信念？

到目前为止，对这些问题的回答还只是模糊的、无法机械化的、片断的，或荒谬的。我们有贝斯理论，登普斯特—谢弗理论，演绎推理法则理论，局部归纳理论，以及大量关于何者最好的争论。但是这些理论最多只能回答上述问题中的一到两个，而且看来没有一个理论是完全正确的。

这种事态虽然没有使我们停止编写医学诊断程序，但是却妨碍了我们对这些程序的理解。没有一个独立的理论可借以确证程序所作的推理。我认为，对一个医学诊断程序来说，如果在临床试验中，病人死得较少，它就比别的程序好。事实上，真正令我烦恼的是，这些程序体现着**不言而喻的**不明推论式理论；只要我们能使这些理论成为显式的，它们就是不明推论式的第一批非平凡形式理论。

看待这种状况的方式有两种：乐观主义的和悲观主义的。悲观主义观点认为：AI 研究者对他们的机会抱着天真的态度，这种态度只是因为对哲学家们昔日的失败一无所知，才得

以维持。我们之所以无法从程序中提取出理论,是因为根本不存在可提取的理论。福多尔(Fodor 1983)在他的《心灵的模化程度》一书末尾不无夸张地总结说:

> 在各类计算的特性中,局部性……是我们知道要加以思考的一个主要特性。让我们看一看……演绎逻辑与证实理论的……对比,前者的发展史无疑是人类探索过程中取得的伟大成就之一;后者[也就是前面所说的不明推论式理论]根据相当普遍的一致意见,很可能是个不存在的领域。我认为,这种不对称的情况……大概并非偶然。演绎逻辑是有效性的逻辑,而有效性是句子的**局部**特性。……一个句子的有效性与它的证实水平截然不同,因为后者……对信念系统的整体特性十分敏感。……直言不讳地说,没有一个计算形式体系告诉我们怎样做到这一点,我们也不知道该如何形成这种形式体系。……在这方面,认知科学甚至尚未起步。说实在的,比起行为主义最黑暗的时期来,我们并没有前进一步。……如果有人,比如德雷福斯,要问我们,为什么我们竟认为在模拟整体过程方面,数字计算机是个似乎合理的机制,那只有以缄默作答了。

当然,根据乐观主义的看法,AI研究者们能够取得比所有哲学家快得多的进步,因为我们有"强大的思想"作武装,而哲学家们没有,特别是复杂的自主计算的思想。我希望这种看法正确。但是如果我们所做的一切只是不断地写程序,而没有任何一般理论出现,我就会变得日益不安。

6. 结　　论

总结：逻辑主义者试图用一阶逻辑表现“朴素物理学”的方案没有大获成功。原因之一可能是其基本论据有缺陷。如果你感兴趣的推理不是演绎，那就无法使写出的公理独立于对它们进行操作的程序。令人遗憾的是，我们感兴趣的推理很少是演绎。同时，逻辑主义者试图扩展逻辑，使其覆盖更多地盘，其结果也令人失望。因此，我们只好老老实实地写程序、把知识表述当作由程序操作的实体。

从多方面来看，这并不是对逻辑**本身**的批评。当你坐下来表达知识的具体内容时，你使用的标记法很快就退到背后去。如果你想建立一个有关形状的理论，那么你的任务就是尽量表达你所知道的一切，而与此无关的由逻辑标记的惯例形成的约束会很快消退。因此，如前所述，我认为莱纳等人(Lenat et al. 1986)的 CYC 方案与逻辑主义方案处在十分相像的阴影下。

然而，逻辑有一个特别易受攻击的方面，这就是它是以指称语义学为基础的。我的结论是：离开程序，形式化的知识就失去效用，对此，人们是能够接受的，但是他们也像我一样能凭直觉强烈地感到：对标记法来说，有指称语义学比没有好。原因之一也许是，健全的语义学至少有助于确保程序作出正确的演绎推理。这些推理也许是微不足道的，但是至少没有错误。

肖厄姆(Shoham 1986)最近指出了另外一种确证形式语

义学的方法。假定程序采用了一个标记法,那么你就可以证明,程序所得的结论正是在其前提的所有A模型中为真的那些结论,至于A模型是什么,则取决于你想获得的推理类别。如果A模型的特性描述在直觉上是有吸引力的,那么你就已经为程序运作的正确性提供了一个独立的证明。如果把"极小模型"嵌入这个构架,我们就得到了一个按一定方式由划界法确证的程序,**只是我们省却了划界公理**,直接使用了语义概念。在时态推理的情况下,我们所需要的模型概念有所不同;参阅肖厄姆(Shoham 1986)的一个提案。这个思想适用于大量不同的推理类型吗?若然,它就提供了一个确证逻辑主义方案的本体部分和语义部分的方法,同时,也罢,免除了非程序的知识表述这一思想。

作为研究知识表述的语义学和构成法方面的问题的工具,逻辑仍然被看作是不可超越的。我想到一些例子,如莫尔(Moore 1980, 1985)对欣蒂卡知识逻辑的计算形式的研究,它解释了思想者怎样才能涉及尚未识别的实体,这些实体的特性别人是知道的;还有查尼亚克(Charniak 1986)把"手稿变量"解释为司寇伦术语的工作。这些文章提出的深刻见解可应用于多种推理程序。那些因为它们用逻辑术语表示的而忽视它们的人,都面临着写出不规范、不切题的程序的危险。

最后我必须承认:我目前仍在做的工作,也属于我这里批评的范式。对于形式表述这个领域,已知的实在太少,所以集中考虑理想化,必然教会我们一些东西。我想要解决的问题是,如何表述为回答像"一个回形针可以当钥匙圈用吗?"这样的问题所需要的知识。我们被迫回到理想化,目的是证明一定尺寸和形状的回形针可以穿过标准的钥匙眼。这样做显然

跳过了原始问题中的大部分。然而,这是一片尚未开发的领地,游览者在穿越理想化地段时,有可能发现某些有趣的东西。在设计出完成一个任务的程序之前,把要了解的关于以不平常方式运用形状的每一件事情都表达为逻辑公理,这是人们无望做到的。这也许会对我和一些朋友造成冲击,但对其他人不会。①

参考书目

Allen, J. (1984). 'Towards a General Theory of Action and Time.' *Artificial Intelligence* 23(2): 123–54.

Birnbaum, L. (1986). 'Integrated Processing in Planning and Understanding.' Yale Computer Science Technical Report 489. New Haven, Conn.: Yale University.

Brownston, L., Farrell, R., Kant, E., and Martin, N. (1985). *Programming Expert Systems in OPS5: An Introduction to Rule-based Programming*. Reading, Mass.: Addison-Wesley.

Charniak, E. (1986). 'Motivation Analysis, Abductive Unification, and Nonmonotonic Equality.' *Artificial Intelligence* [34(1988): 275–96].

—— and McDermott, D. (1985). *Introduction to Artificial Intelligence*. Reading, Mass.: Addison-Wesley.

Clark, K. L. (1978). 'Negation as Failure.' In H. Gallaire and J. Minker (eds.), *Logic and Databases*, pp. 293–322. New York: Plenum Press.

Clocksin, W., and Mellish, C. (1981). *Programming in Prolog*. Berlin: Springer-Verlag.

Davis, R., and Lenat, D. B. (1982). *Knowledge-based Systems in Artificial Intelligence*. New York: McGraw-Hill.

de Kleer, J., and Brown, J. S. (1985). 'A Qualitative Physics Based on Confluences.' In J. Hobbs and R. C. Moore (eds.), *Formal Theories of the Commonsense World*, pp. 231–72. Norwood, NJ: Ablex.

Doyle, J. (1979). 'A Truth Maintenance System.' *Artificial Intelligence* 12(3): 231–72.

---

① 我感谢L·伯恩鲍姆、S·汉克斯、P·海斯、Y·肖厄姆和其他许多人的帮助,有些地方有违他们的本意。我还应当指出,C·休伊特、M·明斯基和B·伍兹长期以来也一直在表述类似的思想。本文得到国际科学基金资助,编号:DCR—8407077。

Duda, R. O., Gaschnig, J. G., and Hart, P. E. (1980). 'Model Design in the Prospector Consultant System for Mineral Exploration.' In D. Michie (ed.), *Expert Systems in the Micro-Electronic Age*, pp. 153–67. Edinburgh: Edinburgh University Press.
Fodor, J. (1983). *The Modularity of Mind.* Cambridge, Mass.: MIT Press/Bradford Books.
Forbus, K. (1984). 'Qualitative Process Theory.' *Artificial Intelligence* 24: 85–168.
Genesereth, M. R. (1983). 'An Overview of Meta-level Architecture.' *Proc. Nat. Conf. AI*, pp. 119–24. Washington, DC.
Hanks, S., and McDermott, D. (1985). 'Temporal Reasoning and Default Logics.' Computer Science Department Technical Report 430. New Haven, Conn.: Yale University.
—— (1986). 'Default Reasoning and Temporal Logics.' *Proc. Nat. Conf. AI*, pp. 328–33. Philadelphia.
Hayes, P. J. (1985*a*). 'The Second Naïve Physics Manifesto.' In J. Hobbs and R. C. Moore (eds.), *Formal Theories of the Commonsense World*, pp. 1–20. Norwood, NJ: Ablex.
—— (1985*b*). 'The Ontology of Liquids.' In J. Hobbs and R. C. Moore (eds.), *Formal Theories of the Commonsense World*, pp. 71–107. Norwood, NJ: Ablex.
Hempel, C. G. (1965). *Aspects of Scientific Explanation.* New York: Free Press.
—— and Oppenheim, P. (1948). 'Studies in the Logic of Explanation.' *Philosophy of Science* 15: 135–75.
Hewitt, C. (1969). 'PLANNER: A Language for Proving Theorems in Robots.' *Proc. 1st IJCAI Conference*, pp. 295–301. Washington, DC.
Hobbs, J. (1986). 'Commonsense Summer: Final Report.' AI Center, SRI International, Technical Note. Menlo Park, Calif.
—— and Moore, R. C. (eds.) (1985). *Formal Theories of the Commonsense World.* Norwood, NJ: Ablex.
Israel, D. (1980). 'What's Wrong with Non-Monotonic Logic?' *Proc. Nat. Conf. AI*, pp. 99–101. Stanford, Calif.
Kuipers, B. (1985). 'The Limits of Qualitative Simulation.' *Proc. IJCAI Conference*, pp. 128–36. Los Angeles, Calif.
Lenat, D. B. (1982). 'AM: Discovery in Mathematics as Heuristic Search.' In R. Davis and D. B. Lenat (eds.), *Knowledge-based Systems in Artificial Intelligence*, pp. 1–225. New York: McGraw-Hill.
Lenat, D. B., Prakash, M., and Shepherd, M. (1986). 'CYC: Using Commonsense Knowledge to Overcome Brittleness and Knowledge Acquisition Bottlenecks.' *AI Magazine* 6(4): 65–85.
Lifschitz, V. (1985). 'Computing Circumscription.' *Proc. IJCAI Conference*, pp. 121–27. Los Angeles, Calif.
Luckham, D. C., and Nilsson, N. J. (1971). 'Extracting Information from Resolution Proof Trees.' *Artificial Intelligence* 2(1): 27–54.
Marcus, M. P. (1980). *A Theory of Syntactic Recognition for Natural Language.* Cambridge, Mass.: MIT Press.
Marr, D. (1977). 'Artificial Intelligence—A Personal View.' *Artificial Intelligence* 9: 37–48.
McCarthy, J. (1980). 'Circumscription: A Nonmonotonic Inference Rule.' *Artificial Intelligence* 13: 27–40.
McDermott, D. (1978). 'Tarskian Semantics or, No Notation Without Denotation!' *Cognitive Science* 2(3): 277–82.
—— (1982). 'A Temporal Logic for Reasoning about Processes and Plans.' *Cognitive Science* 6: 101–55.
—— (1985). 'Reasoning about Plans.' In J. Hobbs and R. C. Moore (eds.), *Formal*

Theories of the Commonsense World, pp. 269–317. Norwood, NJ: Ablex.
—— and Doyle, J. (1980). 'Non-Monotonic Logic I.' *Artificial Intelligence* 13: 41–72.
Moore, R. C. (1980). 'Reasoning about Knowledge and Action.' AI Center Technical Report 191, SRI International. Menlo Park, Calif.
—— (1985). 'A Formal Theory of Knowledge and Action.' In J. Hobbs and R. C. Moore (eds.), *Formal Theories of the Commonsense World*, pp. 319–58. Norwood, NJ: Ablex.
Newell, A. (1981). 'The Knowledge Level.' *AI Magazine* 1(3): 1–20.
Reiter, R. (1980). 'A Logic for Default Reasoning.' *Artificial Intelligence* 13: 81–132.
Rosenschein, S. J. (1981). 'Plan Synthesis: A Logical Perspective.' *Proc. IJCAI Conference*, pp. 331–7. Vancouver, BC.
Salmon, W. C. (1967). *The Foundation of Scientific Inference*. Pittsburgh: University of Pittsburgh Press.
—— (1975). 'Theoretical Explanation.' In S. Koerner (ed.), *Explanation*, pp. 118–45. New Haven, Conn.: Yale University Press.
Shoham, Y. (1985). 'Naïve Kinematics: One Aspect of Shape.' *Proc. IJCAI Conference*, pp. 436–42. Los Angeles, Calif.
—— (1986). 'Time and Causality from the Standpoint of Artificial Intelligence.' Ph.D. diss. Yale University.
—— and McDermott, D. (1984). 'Knowledge Inversion.' *Proc. Nat. Conf. AI*, pp. 295–9. Austin, Tex.
Shortliffe, E. (1976). *Computer-based Medical Consultations: MYCIN*. New York: Elsevier.

10 动机、机制和情感

A·斯洛曼*

1. 引　　言

日常语言在不同种类的心理状态和心理过程，例如情绪、情感、态度、动机、性格、个性等等之间造成了丰富而微妙的差别。在实际需要的驱动下，我们的词汇和概念经过数世纪错综复杂的实际生活的磨砺，从而深刻地映射出人类的心灵。

然而实际使用起来却有不一致之处，对于那些我们所掌握的、并能在直觉上运用的差别，我们能清晰地加以描述的能力是很有限的，就和我们背诵英语句法规则的能力一样。像"动机"和"情感"这些词，在使用中存在着许多含混不清和不一致的地方。同一个人会对你说，爱是一种情感，她深深地爱着她的孩子们，同时她又不处在情感状态之中。如果我们用审慎定义的术语重新陈述这些说法，就能够对许多不一致性作出解释。作为科学家，我们需要用具有理论基础的术语来补充口头语言，这些术语可用来指明差别和描述那些一般人通常难以觉察的种种可能的情况。例如，根据这种看法，爱是

一种态度，而不是一种情感，尽管挚爱很容易引发情感状态。用哲学家的行话来说(Ryle 1949)，态度是倾向，情感虽然带有倾向的成份，却是偶发的情节。

为了充分说明这些情节和倾向，我们需要一个关于怎样产生和控制心理状态、以及它们怎样导致行动的理论——一个关于心灵机制的理论。该理论应当解释内部表述是怎样被建立、存储和比较的，又是怎样用来作出推理、制定计划或控制行动的。本文将提出一个理论纲领，概述适用于智能动物或机器的种种设计约束，然后将设计解答与人类动机结构以及与作为常见情感状态基础的计算机制联系起来。

经分析，情感可表现为多种状态，在每一状态中，在资源有限的智能系统所必需的引发机制的作用下，强有力的动机对有重要意义的信念作出响应。新的思想和动机设法通过各式各样的过滤器，造成对另一些正在进行的活动的干扰。其结果有可能打断或修正别的心理和物理过程的运作，有时是富有成效的，有时则否。这是一些被"移动"的状态。生理变化不一定参与其中。情感与感觉、冲动、情绪、态度、气质这样一些相关的状态和过程有一些微妙的差别；但因篇幅所限，这里不能详细讨论。

根据这一观点，既然以智能为基础的机制已能满足要求，我们就无需用专门的子系统来说明情感(参阅 Oatley and Johnson-Laird 1985)。如果因为在复杂快变的世界中需要有以智能方式行事的机制，而造成情感状态的出现，那么通常认为

---

* A·斯洛曼的"动机、机制和情感"原载《认知和情感》1(1987)：第217—233页。作者和 Lawrence Erlbaum 联合有限公司允许重印。

A·斯洛曼(Aaron Sloman)，苏塞克斯大学认知与计算科学学院人工智能教授。

情感和认知相分离的看法就受到挑战。这同样适用于人类、其他动物或未来的智能机器。

2. 心灵的设计约束

多种多样的动物行为表明,存在着有心计的行为者的各种不同的行为方式,这些行为者吸取来自环境的信息,并能根据信息独立地或联合地采取行动。人类仅仅占据了这种"可能心灵空间"的一个角落。我在别处简略讲过一些约束,它们决定着体现在人类心灵中的设计解答,这里我只能对与情感有关的重要结果作一概要论述。

各种约束包括:来自内部和外部的动机源的(常常是不一致的)多重性、速度的限制、对环境看法的难以避免的空缺和错误、与动机相关联的变动着的紧迫程度。由于资源的限制和紧迫性,不可避免地要使用潜在地不可靠的"快速估算"策略。新信息和新目标的不可预测性意味着需要具有打断、修正、延缓或中止当前活动的能力,不管这些活动是内部的还是外部的。这里包括诸如硬件上的和软件上的"反射"行动之类的东西,其中有些需要根据经验作出修正。

反射是以内在方式作快速行动,但又是愚笨的。它们可能部分受控于使用"快速估算"的背景敏感过滤器,从而迅速对优先权作出估计,并使正在进行的极端重要的、紧急的或危险的活动免遭干扰而继续进行,同时又允许新的、特别重要的或紧急的动机打断这些活动。(后面将根据动机通过这种过滤器的能力,来定义动机的"坚持性"。)一个重要的结论是,智

能系统，包括用于新动机的过滤器，具有快速但愚笨的子系统。快速而愚钝的过滤器时常会招致令人不快的后果。

由于信息不完全，而且要处理社会或物理环境中的长期变化，就需要较高级的可供学习之用的行动源：不仅有动机生成器和比较器，还有用于生成器和比较器本身的生成器和比较器。

虽然数个独立的子系统可以并行地执行各自的计划，如边走边吃，但是需求方面的冲突会产生不相容的目标，所以必须有一个决策制定机制。有两种主要方式可供选择："民主"投票式，中央决策者式。如果不是所有子系统都能接近全部已知信息存储，或者它们不是都有同等的推理能力，那么"民主式"组织形式就可能是危险的。另一方面，在作出重要决策时，需要有专门化的中央机制（Sloman 1978：chs.6 and 10）。看来，这就是正常人类心灵具有的组织形式。

类似的约束对智能式人工产品的设计具有决定性影响。由于生物或人工计算装置受到物理方面的限制，所以必须对各种功能作出主次分工，包括将最高层次的控制权分配给能获取最多信息和具有最强有力推理机制的部分。然而，在偶然的对强烈行动的紧急需求之下，必须具备压倒性的硬件和软件上的反射，这些反射不依靠较高层次的控制权而独立运作——这一机制有可能使下文中所说的情感过程得以实现。

目标生成器

许多不同种类的动机激发因素都是由日常词汇和短语表明的，如：

目的、态度、欲望、厌恶、目标、憎恨、希望、理想、冲动、喜欢、爱情、爱好、原则、诱惑、娱乐、悲伤、厌烦、迷惑、高兴、沮丧、痛苦。

此外还有很多。它们反映出在不同行动源和影响我们的事物的各种方式之间所存在的微妙差别。概念分析(Sloman 1978: ch.4)为它们建立了预设假定。一个关键的概念就是持有目标。

取初步近似,持有目标就是以某种形式结构表述的符号结构来描述有待产生、保存或防止的事态。符号不一定非得是物理结构:虚拟形式体系也可胜任(Shoman 1984)。目标能够使用与信念和假设完全相同的描述性形式体系,区别仅在于它们所起的作用。

事态的表述可以起到目标的作用,只要它趋于(在许多先决条件制约之下)产生改变现实使之与表述内容保持一致的那种行为。

一个表述可以起到信念的作用,只要那些产生或修正它的感觉和推理的过程趋于(也在许多先决条件制约之下)改变表述使之与现实保持一致。

(这里"保持一致"的意思须得长篇离题之论才解释得清楚。)同样的表述也可用在其他方面,如指令、所假设的情境、规则等等。

有一些新目标有利于前目标的实现,它们是由计划过程

生成的。有一些则是对新信息的响应，比如想了解一下拐角那边十分吵闹的原因是什么。目标不仅仅是通过外部事件引发的，在这方面，思想、推理或回忆很可能有同样的效果。

信念或思想怎样能产生目标呢？如果目标含有符号结构，计算解释可能就是采用“生成目标”的条件-行动规则。例如，慈善规则或许是：“如果X处在痛苦中，就生成一个目标[解除X的痛苦]。”一个处在发怒状态下的惩罚目标生成器或许是：“如果X伤害了我，就生成一个目标[使X受到惩罚]。”详尽的分析可以描述各种“目标生成器”、“目标生成器的生成器”，等等。根据经验，学习系统会使用生成器的生成器来产生新的目标生成器。

目标比较器

生成器产生的目标并不总是协调一致的。不同的设计约束导致不同的目标生成器共存。社会性动物或机器需要一些产生有利于其他动物或机器的目标的目标生成器，这些目标有可能与各自的目标和需要相冲突，因而需要有目标比较器对不同的后果作出选择。

有些比较器在计划中采用约束目标，例如使用“极小代价”规则，在两个子目标中选出代价较小的那一个。另一些则命令结果直接出现，像拯救生命规则，认为与其他任何目标相比，拯救生命总是更重要的，而不是运用某种通常方式对两者作出裁定。因为引起动机的根源是不同的、不可通约的，而且比较的基础也不同，所以对于一个冲突情况不一定有**最优**解答。

较高级的动机激发因素

尽管口语的内涵有可能混淆，术语“动机激发因素”一般用来指根据信念趋于产生、修正或选择行动的机制和表述。动机激发因素以递归方式包含着动机激发因素的生成器和比较器。某些动机激发因素是暂时的，例如获得一块特制蛋糕的目标，而另一些则是长期的，例如希望保持身材苗条。

动机激发因素不应该是静态的，要灵活地产生新目标，就需要动机激发因素生成器。更高程度的智能还具有从经验中学习和修正生成器的能力。因此就要求生成器是递归的。这同样适用于比较器：如果两个生成器由于生成相冲突的目标而经常出现冲突，那么就可能需要废止或修改其中的一个。这样就需要有生成器的比较器。同时也需要比较器的生成器和比较器的比较器。较高级生成器和比较器可说明某些个性差异。它们的作用可用来说明情感状态的某些微妙之处。

为了设计出普遍适用的较高层次的生成器和比较器，需要进行理论研究。为了弄清楚人类所具有的机制，也需要经验研究。对于生成器和比较器的层次，我们是否有一个极限，或者说，新的层次是否能无限制地递归生成？

多种多样的动机激发因素

我们可以对“派生”和“非派生”的动机激发因素加以区分。粗略地说，如果一个动机激发因素明确地是从另一个动机激发因素通过“手段目的”分析而得出的，并且这一来源被记录下来，在后继的处理中发挥作用，那么所得出的动机激发

因素就是派生的。口渴时想要喝水的愿望是非派生的，而为买饮料想要钱的愿望则是派生的。一个动机可以部分是派生的，部分是非派生的，比如为了使别人印象深刻，而用威士忌止渴。下面我们尝试说明非派生的动机激发因素是怎样对情感状态起核心作用的。

这种差别与行为有关。派生目标比较容易被放弃，放弃它们的副作用比较小，例如，当它们看起来无法达到时，或者当产生它们的那些目标已被满足或放弃时。一个没有希望的派生目标很容易被另一个所取代，如果后者也能使上层目标得到满足的话。如果非派生目标由于与另一些被认为是更重要的目标发生冲突而被放弃，它有可能继续要求得到注意，形成一个情感源。如果能将不相容性去掉，放弃的做法会带来懊悔，并产生恢复这一目标的倾向。

人类的非派生目标有：身体的需要、得到赞同的愿望、好奇心、对美的追求，以及承担任务时获得成功的愿望。由于这些目标是为更一般的生物学目的服务的，有些理论家把它们看成派生的。然而，目标的生成机制不必明确地把这目标与一些较高层次的目标联系起来，而只是赋予这目标以产生计划和行动的因果能力，例如通过直接在数据库中插入目标表述，该数据库的内容则不断地驱动这一系统。尽管这样一个目标的作用是隐含的，它对个体来说仍是非派生目标。

变动情况的定量维度

动机可以在不同的维度上进行比较，这些维度可根据前面概述的机制来定义。动机的**坚持性**表现为它的中断性优

先权水平。被坚持的欲望、疼痛、恐惧等等，是一些更容易通过中断过滤器的动机，它们取决于当前活动而设置的阈值。

必须对通过过滤器的目标进行比较，以评价它们的相对**重要性**。这种(有时是部分的)次序关系是由信念和比较器决定的，如果信念和比较器发生变化，次序关系也会发生变化。也可能需要复杂的推理。派生目标的重要性是与对达到或未达到该目标所产生的结果的信念相联系的。坚持性关系到一个目标有可能怎样通过中断过滤器以便被考虑；而重要性则关系到在被考虑的情况下，被采纳为应当做的事情的可能性如何。对坚持性必须非常迅速地作出评价，并应与重要性联合考虑，但有时并不如此。一个劣等过滤器会把低度的优先权分配给重要目标，或是把高度的优先权分配给不重要的目标。打喷嚏的欲望并不因为安静是生存的要素而消失。(并不是所有动物都有这样复杂的动机激发系统。)

紧急性是对多长时间内事情不至于被延误的量度。这不同于坚持性或重要性：有些并非渴望的事情可能是紧急的，反之亦然。

目标的**强度**决定着一个目标被采纳之后，以何种积极或强烈的方式来实现它。它部分地与紧急性和重要性有关，又部分地独立于它们。强目标遇到的障碍往往被当作一种挑战，而不当作放弃目标的理由。一个重要的长期目标，常常败在某些重要程度低得多，但却更强的事情的手下，这种欲望与职责之间的冲突由来已久。在理想状态下，坚持性、强度和重要性应是相互关联的，但是这种关系可能因与紧急性的相互作用而被打乱，以及因从先前的经验或进化的起源产生的坚持性或强度的反射分配方式而被打乱。

还有另一种度量动机的方式：如果它不能实现，痛苦或破坏程度如何。而如果动机实现，获得喜悦的程度也是不同的。要对此作出评价，可以估算使这种满足状态得以保持，或是在以后再次获得它，需要付出多大努力。这两者通常可用某人对此“关心”的程度是多大来表示，同时也与下文提到的生成情感状态的潜能有关。

动机的这些不同种类的“力量”，在实现认知功能的过程中全都在起作用，它们可能也是复杂的机器人所需要的。在一个有自监控的系统中，它们可能具有主体式的相互关联，尽管自监控并不总是完全可靠的。客观地说，它们是通过产生各种不同的内部或外部结果，或抵制各种不同的内部或外部变化的倾向来定义的。各种力量的不同组合，将影响着从初始构想到实现、放弃或失败的各个目标演进阶段上会发生的事情。

3．与动机有关的过程概述

至此，我们已经略述了动机可能经过的下列中间过程：

1．始发——由身体监控器、动机生成器、或建立新的子目标的计划器所产生。

2．新目标的反射性优先——对坚持性的分配。

3．抑制或通过——由中断过滤器确定。

4．引发反射行动（内部的或外部的，硬件或软件）。

5．评价相对重要性——由比较器进行。

6．采纳、排斥或延迟考虑——已被采纳的动机一般称为

"意向"。欲望如果没有变为行动,仍可作为欲望存在。

7. 制定计划——"内在"计划与怎样实现目标有关,"外在"计划与什么时候、怎样使它与其他活动相联系有关,例如,它应该推迟吗?

8. 激活——开始实现动机,或是重新激活暂时搁置的动机。

9. 计划的执行。

10. 中断——放弃或搁置。

11. 与新目标的比较。

12. 计划或行动的修正——根据新信息或新目标,包括速度、式样或子目标的改变作出。

13. 满足(全部或部分)。

14. 挫折或妨碍。

15. 内部监控(有自我意识)。

16. 学习——根据经验对生成器和比较器进行修正。

这些都是计算过程,可通过由规则支配的对各种表述的操作来表示,虽然详尽地描述这些细节不是一项轻而易举的任务。下面我打算说明它们怎样与情感相联系。这整个情况是非常复杂的。

例示:发怒。"X 对 Y 发怒"的意思是什么?这表明,X 相信 Y 做了某件事,或是没能做某件事,致使 X 的动机之一受到妨碍。这一信念和动机的结合并不足以造成发怒,因为 X 也许只是对所发生的事情感到遗憾,或是对 Y 感到失望,而没有发怒。要发怒,X 还需要有一个新动机:一个伤害或损害 Y 的欲望。大多数人和许多动物似乎都有报复动机生成器,它作出的反应就是这样的,可叹!在行动中,不是必然会选择

新动机,尽管该动机可能是很强烈的:对后果的担心以及合适的比较器可能使它无效。

产生了新欲望还不足以引起发怒。很可能 X 虽然有这个欲望,却没有把它放在心上,而是平静地继续做另一些事情:在这种情况下,他是不会发怒的。也有另一种可能,X 要做使 Y 不愉快的事情的欲望可能完全是派生的,纯粹是从减少将来出现这种事情的可能性这一实际效果出发的,并非对 Y 怀有恶意。所以如果 X 能够以某种方式确知这一事件不会再次发生,就会放弃这一动机,不再发怒了。

发怒包含着**坚持**而**强烈**的要做些事使 Y 蒙受痛苦的非派生欲望。高坚持性意味着该欲望频繁地通过 X 的过滤器,“提请”X 的决策制定过程“注意”。所以即使比较器对此作出否定,该欲望仍频繁地回到 X 的思想中,使他难以把精力集中在其他活动上。为速度设计的过滤器可能太愚钝,不足以排除已被较高层次否定的动机。此外,该欲望必须不是派生的,派生的子目标会因上层目标的取消而消失。在复杂的社会动因中,发怒可能包含着认为 Y 的行动不符合社会或伦理标准的信念。

所以情感是由动机激发因素产生的状态,同时包含着新动机激发因素的产生。

原有动机受到妨碍以及新动机得到坚持,可能与附加的次级作用相联系。例如,如果 X 得知自己发怒了,可能引起他的烦恼。如果其他人觉察到他的状况,也有可能对这一情感的性质施加影响。这一情节可能唤起对另一些增强发怒情绪的局面的回忆。

有时候,人类的情感状态也会产生生理上的波动,正如前

面提到的那样，这可能是受“快速估算”策略驱使的物理和化学反射运作所造成的。然而，如果 X 充分满足了其他条件，他就可以正确地被说成是怒气很大，即使没有任何身体征兆。强烈的发怒也可能在没有任何身体副作用的情况下存在，只要它不断闯入 X 的思想和决策，只要 X 强烈地希望给 Y 造成痛苦，而且是巨大的痛苦。虽然非体征的发怒可以称为“冷酷”，但从社会角度看，它依然具有发怒的全部重要方面。

发怒**实际上**不一定会干扰其他动机，因此它只是一种局部倾向：例如，如果惩罚 Y 的新动机付诸行动了，就不必再做进一步的干扰。然而，发怒具有扰乱其他活动的**潜能**，如果新动机具有高坚持性的话。

由于自我监控的作用，有时会**感觉到**在发怒。然而也可能处在发怒或别的情感状态中而没有意识到它的发生。例如，我怀疑狗和幼童是不知道他们在发怒的（虽然他们十分了解是什么激起了发怒）。

像发怒这样的情感可以在不同的定量维度和定性维度上做变动，诸如：X 对于 Y 所做的事情确信到什么程度，X 对此关心到什么程度（即：受妨碍的动机的重要性和强度如何）；X 希望让 Y 受到多大的伤害；这个新欲望的重要性如何，强度如何，坚持性如何，能延续多长时间；对 X 产生的心理干扰有多大；生理干扰有多大；X 意识到这一状态的哪些方面；会生成多少次级动机和行动。不同的维度适用于不同的情感。

在事情的各个不同阶段上的变动对应于不同的状态，其中有些不是情感状态。如果不存在要伤害 Y 的欲望，这一情感更像是愤怒，而不是发怒。如果根本没有响应的属性，这一情感只不过是某种形式的烦恼，而如果受妨碍的动机非常重

要，同时又不能立即用别的替代物来满足，那么这一情感就包含了沮丧。由于任意多的动机、信念和动机生成器可以被包含进来，同时旧反应的作用会引发新反应，所以这一理论所覆盖的变动范围必然比日常语言中的标号集更加丰富，也比生理响应的范围更加丰富。

4. 有关对情感的生成语法

根据上述机制对发怒和其他情感进行的分析，表明情感状态是由以下成分组成的：

1. 至少有一个初始发动机 M1，处在重要性和强度的高水平上。

2. M1 得到满足或受到妨碍，无论是真实的、想象的或预期的，对此所持的信念 B1 都会引发不同种类的生成器，常常产生新的动机。

3. 各种不同的状况取决于：(a)M1 是与所想望的还是与所厌恶的某件事有关；(b)B1 是关于 M1 得到满足还是受到妨碍的信念；(c)B1 同过去、现在还是将来有关；(d)B1 是否包含着不确定性；(e)行为者是否意识到自己的情感；(f)是否认为涉及其他行为者；(g)M1 是否与其他行为者对该行为者的看法有关（参阅 Roseman 1979）。

4. 在更复杂的情境中，数个动机同时与信念相互作用，如在一个情境中，B1 示意重要动机 M1(a)与 M1(b)是不相容的，又如，处于进退两难的时候。

5. 有时 M1 和 B1 引发了一个产生次级动机 M2 的生成

器，例如想把事情做好、保持愉悦、惩处罪犯或将消息告诉别人的欲望。这又转而能同另一些信念相互作用，对认知过程造成干扰、中断，或产生别的影响。这将是一种“两层次”的情感状态。也可能出现数个层次。

6. 有时 M1 和 B1 同时引发数个动机生成器，造成的相互作用可能是非常复杂的，尤其是当几个新动机处于冲突状态时，比如一个既要避免破坏，又要捉拿罪犯的欲望。

7. 有时，新生成的动机与以前存在的动机发生冲突。

8. 高度坚持的新动机通过中断过滤器，往往会产生（虽然它们实际上未必产生）一个干扰，即不断地打断思维和决策，并影响着制定决策的判据和感知能力。

9. 思想和动机都可能中断。即使没有新动机，也可能简单地存在着在 M1 和 B1 上的持久滞留。像悲痛这样的情感尤其是如此，其中涉及一些无法解脱的事情。这种被迫性强调可能来自有关重编程生成器的自动学习机制的引发。

10. 不必为行动选择新动机。M2 有可能被看作不重要的而受到拒绝，然而如果它高度坚持，就会继续通过中断过滤器。

11. 如果处于某些情感状态，如恐惧，M2 会绕过仔细思考和做计划，并打断其他行动，而引发反射行动（Sloman 1978: ch.6）。“软件式反射”被称作“冲动”行动。由于反射，有可能采取非常迅速的补救行动，或抓住突然的机会。然而，有时它们也会是灾难性的。有些反射纯粹是心理的：接二连三的思想和感情会全都被引发出来。

12. 有些情感状态起因于个体本身的思想或行动，例如，因预期可能出现错误而产生担心。为了过分的关心等等，可

能生成次级动机。这些次级动机可能生成大量干扰,从而导致灾难性后果。

13. 有些情感会使许多正在进行的过程中断,并改变其方向,例如在恢复平衡时控制身体各部分的那些过程。如果感觉探测器记录了局部的变化,系统对自身状态的感知就会改变。

14. 自监控过程可能探测到新的内部状态,也可能探测不到。如果探测不到,X 就无法意识到或感觉到情感。内部监控不一定产生识别,例如,也许还未学习过有关的图式(Sloman 1978:ch.10)。人们必须学会区别和识别复杂的内部状态,这里使用的感知过程的复杂性决不小于面孔或打字机的识别过程。

15. 对情感的识别能够产生进一步的影响,例如,在内部状态实现了或是妨碍了某个动机的时候。它有可能激活休止的动机或动机生成器,也可能导致逐步走向更高级的情感(递归式上升)。

作为某些情感的特征的中断、干扰和对理性的背离,是各种机制自然出现的结果,这些机制产生于为智能系统设计的各种约束,特别是必须快速行动而资源有限的中断过滤器的不可避免的愚钝性。一个有无限快的计算机并具有完备知识和预见能力的机器人,也许就不需要这些机制了。然而,并非所有情感都表现出机能障碍:在狭窄的山脊上行走时,不忘记种种危险,对你来说是很重要的。

这些机制可使不同情境中如此之多的不同子过程存在,所以不能列一个关于情感类型的简单的表以准确合理地反映出这些变化。聪明的、有自监控能力的机器人身上的细微的

情感现象，可由同样的丰富变化来刻划。

要对人类觉察发怒、得意、恐惧等等的一般方式作出充分说明，势必涉及身体感受。然而在我们的生活中，许多情感之所以重要，并不是由于这种细节，而是由于更具整体特征的认知结构。狂怒之所以被重视，是因为它会产生那些对仇恨者和被仇恨者造成伤害的行动，而不是因为身体紧张和大汗淋漓。悲痛之所以值得注意，是因为失去了心爱的孩子，而不是腹中有一种新的感觉。所以我们用像"害怕"、"失望"、"大喜"、"狂怒"或"极度悲伤"这样一些术语，来描述一个异己者甚或没有生理反应的、高度精密的机器人的心态，是合理的(可对照 Lyons 1980)。

5. 情绪、态度和个性

情绪与情感有相像之处，它也对心理过程造成某种整体的干扰，或是具有干扰倾向。然而，它不必包含任何专门的信念、欲望、行动倾向等等。就人类来说，情绪可以由化学因素或认知因素，例如由饮酒或听到好消息或坏消息所诱发。情绪可以使人们感知事物、解释他人行为、预见行为结果、制定计划等等的方式具有色彩。和情感状态一样，情绪可以被、也可以不被有关个体感知和分类。更详尽的理论必须区分不同的机制，例如，某些子过程中由"硬件引起的"整体速度变化，和在动机的相对优先权中或推理策略中由"软件引起的"整体变化。

态度，如爱或钦佩，是集中于某一个人、物体或观念的信

念、动机、动机生成器和比较器的集合。爱自己孩子的人，在发觉有危险，或看到可能影响他们幸福的情况时，就会争取新的目标。爱的力量决定着分配到这些目标上的重要性和中断性优先权水平。自私是对待自我的一种类似的态度。在能够思考和关心他人心理状态的智能系统群体中，态度的丰富多变使它们成为诗人、小说家和社会科学家们用之不竭的研究课题。态度常常与情感混淆，事实上在根本没有情感参与的情况下，也可以有爱、怜悯、钦佩或憎恨。**在有机会的时候**，态度就会在作出某种选择的倾向中表现出来，但是它们不一定包含对思想和决策的无休止的干扰。人们可以并不一直想着自己的孩子，而仍然爱着他们，虽然自己所爱的人有危险的消息会引发情感。

性格和个性包含着长期的态度。例如，慷慨不是一个目标，而是一簇目标生成器和比较器，生成器产生出回应有关别人需求的信息的新目标，比较器则选择新目标而不选择更加以自我为中心的目标。伪善者也产生类似的目标，但从不将它们付诸行动。个性或性格汇集着为数众多的、在特殊场合产生特定目标的、没有特定指向的一般性倾向。这样的汇集体集合的丰富性是日常语言难以形容的。复杂的个性可能要用整部小说来刻划。更为常见的是，可能的心理状态和过程的空间过于丰富和复杂，使诸如"态度"、"情感"、"情绪"这些口语化的说法无法用于合适的科学理论。

由于缺少更丰富、更细致的词汇，我们把多种深刻而感人的经验说成是情感，如观赏美景，阅读诗歌，听音乐，专注于一部电影或一个问题。这些经验涉及感知与大量附加过程之间的强有力的相互作用，这些过程既有身体方面的，也有精神方

面的。听音乐可以引起身体的运动,同时还产生大量的心理"运动":记忆、感觉、联想的波动——所有这些都在音乐的统摄之下。对这些过程的解释,可以借助于这里未曾讨论的智能系统设计的种种方面,诸如对联想记忆的需要,控制身体运动时进行整合和同步化的微妙方式。在个体内部以及在从事于协作任务的个体之间,这种同步化都是不可缺少的。音乐似乎驾驭了这样一些过程。

我推测,本文概述的这些机制能够生成我们平常所说的那些情感状态——恐惧、发怒、失意、兴奋、沮丧、悲伤、愉悦等等。与之有关的动机、信念、计划和社会背景的复杂程度可能是无止境的,而它们生成的情感过程也可能同样复杂多变,从这个意义上说,这些机制是生成的(Abelson 1973; Dyer 1981; Lehnert, Black and Reiser 1981)。这就是说,对情感状态来说,没有一种简单、有限的分类法能够着手去掌握这种多变性,正像英语句子的多变性也不能由一种分类法来掌握一样(参阅 Roseman 1979)。

心灵科学理论需要这样的概念吗?

常常有这种看法:虽然像"信念"、"欲望"、"情感"这些概念在个体对他人的认识中起着重要作用,但是形成完备的心灵科学理论,并不需要这些概念。这种看法的极端形式是唯物主义的还原论。但是对于心理学来说,这是难以置信的,这就如同说软件设计的概念可以用那些仅仅与计算硬件有关的概念代替一样。

一个更加精致的提法(S. Rosenschein, SRI, 个人通信)是,

与信念、欲望、意向等无关的“中间层次”的概念组成了一个全新的集合，这一集合足以使科学理论成功地预见和解释人类和其他智能生物是怎样工作的。看来在对人类行为作出有重要意义的概括时，日常概念不大可能被完全摒弃（Pylyshyn 1986），因此我采取了妥协的立场：我们不是将日常概念全部替换，而是要将其扩充和提炼，从而说明它们怎样同实际的设计规定建立联系。

即使这种理论是错误的，它也可能同涉及人类心理状态和行为的自然语言概念的语义学有着深刻的内在联系。如果是这样，一台能够理解日常语言并能模拟人类的交流方式的机器，至少需要隐含地掌握这一理论。

启　示

这些机制并非都能在所有动物中找到。在某些智能较低的物种中，动机的选择很可能无法与发起行动的过程相分离，操作性动机不可能是潜伏的。这种动物和机器缺乏灵活应付复杂环境所需的那些机制，因而上述意义的情感就不可能出现。

在幼童中，这种丰富性也是不大可能全部存在的。通过对儿童形成认知和计算机制，以及动机激发机制的调查，我们可望对儿童的情感状态有更多的理解。特别应指出的是，看来许多较高级生成器和比较器是婴儿身上没有的，中断过滤器的选择性也远远低于大多数成人，如果软件式过滤器是学习的结果，这种情况就不足为怪了。

正是上述机制的复杂性，揭示了“出毛病”的众多方面。

动机生成器和比较器会产生不利的欲望和偏爱。中断性优先权的分配方式可能没有与对重要性所作的反射判断相互联系。中断阈值可能设置得过高或过低。对生成器和比较器作出修正的学习过程可能太快,无法根据不充分的证据将事物改变。由于某些较快的反射和过滤器的不可避免的愚钝性,可以预料,生成器和比较器的某种功能障碍将导致强烈的情感,对正常的认知或社会活动造成干扰。对未能实现的动机所作的反应,有可能太强或太弱,结果使个人的或他伙伴的长期利益受损。通过目标评价过程,分配给不同种类动机的相对重要性,可能产生一种倾向,就是去选择那些无法实现或需要付出巨大代价才能实现的目标。潜伏的、暂时搁置的动机可能常常被忽视,这是因为监控过程未能觉察到时机,也可能因为检索不当。为应付信息不足、资源有限和速度需求而滥用"快速估算法",会造成相当广泛的系统性的功能障碍。情感的递归上升很可能是造成某种紧张状态的原因。

许多常见类型的错误是难以避免的,那些希望在不太遥远的将来,即能通过机器对重要决策作出快速处理的人,应当对此予以足够的重视。

事实上,如果人类像我们所认为的那样错综复杂的话,那么有如此之多的稳重而有教养的人,就是一件令人吃惊的事了。或许,这一理论将揭示出我们以前不能识别的种种干扰类型。

这一理论表明,由动机激发因素组成的、复杂而频繁变化的集合可作为一个框架,学习和认知发展的过程就是在这一框架中发生的。这些过程以及它们所生成的那些过程,必然会对学到的东西产生深远的影响,人们正期待着这一天的到

来，届时个体之间将会存在巨大的差异。它给教育工作者带来的启示，还有待于探究。

6. 结　论

这种一般性的理论是关于心灵的**计算**理论。这些计算可存在于**虚拟**机中，由较低层次的似脑或似计算机的机器来完成，而不必直接由物理过程来完成。因此，一方面是那些可以在常规计算机中的低层次上见到的以物理方式作出的显式表述，一方面是那些神经网络模型中研究的隐含的或分布式的表述，这一理论处在这两者之间的中性位置上。

对这一方法的检验，将表现出以该方法为基础的理论所具有的解释能力。我们既需要对我们在人类行为中发现的种种可能性作出全范围的、系统的解释，也需要对人类之间以及人类与其他实际的和可能的行为实施系统之间如何加以区别作出说明。（关于可能性的解释，参阅 Sloman 1978：ch.2。）

理解我们所熟悉的心理过程背后的计算机制，可使我们有能力减少由情感扰动造成的痛苦，减少学习能力的缺陷，以及减少社交能力方面的许多不足。有些问题可能是脑损伤或神经功能障碍所致，另一些问题看来更像是计算机中的软件失误。我推测，很多为情感所困扰的人，正在经历这样的软件式“毛病”。

这一分析还遗留下许多空白。特别是没有对愉悦和痛苦作出解释，同时，对于感到某种事情可笑是怎么回事，我也没有作出能为人所接受的分析。有一些状态，如因快速运动而

震颤，因日落而出神，因读一本书或看一出戏而感动得落泪，这些都需要作出更加详细的分析。人类情感活动的许多方面是在人类进化史上偶然出现的，而且不是设计精良的机器人所必备的，对此我都没有论及。因而要做的工作还很多。虽然如此，该理论仍为思考一系列可能存在的自然的和人工的智能系统类型，提供了一个框架，这是我们对可能存在的心灵的空间所作一般性研究的一部分。通过实际的计算机模拟来检验这些想法，必定会使诸多遗留问题和缺点显露出来。[①]

参考书目

Abelson, R. A. (1973). 'The Structure of Belief Systems.' In R. C. Schank and K. M. Colby (eds.), *Computer Models of Thought and Language*, pp. 287–340. San Francisco: Freeman.

Austin, J. L. (1961). 'A Plea for Excuses.' In *Philosophical Papers*. Repr. in A. R. White (ed.), *Philosophy of Action*, pp. 19–42. Oxford: Oxford University Press.

Boden, M. (1972). *Purposive Explanation in Psychology*. Cambridge, Mass.: Harvard University Press.

—— (1977). *Artificial Intelligence and Natural Man*. Hassocks, Sussex: Harvester Press.

Croucher, M. (1985). 'A Computational Approach to Emotions.' Unpublished thesis. University of Sussex.

Dennett, D. C. (1979). *Brainstorms*. Hassocks, Sussex: Harvester Press.

Dyer, M. G. (1981). 'The Role of TAUs in Narratives.' *Proceedings Cognitive Science Conference*, pp. 225–7. Berkeley, Calif.

Edelson, T. (1986). 'Can a System Be Intelligent If It Never Gives a Damn?' *Proc.*

① 该项工作受到 GEC 研究实验室研究基金的资助，并得到复兴信托基金机构的支持。有不少思想来自西蒙的开创性著作。我的观点还得益于数年来同 M·克劳奇的讨论，她的论文（Croucher 1985）使本文中提出的观点大为深化。本文概述的、并在我 1978 年著作（Sloman 1978）的第 4 章中加以扩展的把日常语言看作有关人类心灵的信息来源的观点，从奥斯汀（Austin 1961）那里获益匪浅。K·奥特利的编辑评论也有极大帮助。

5th Nat. Conf. AI (AAAI-86), pp. 298–302. Philadelphia.
Heider, F. (1958). *The Psychology of Interpersonal Relations*. New York: Wiley.
Lehnert, W. G., Black, J. B., and Reiser, B. J. (1981). 'Summarising Narratives.' *Proc. 7th IJCAI Conference*, pp. 184–9. Vancouver, BC.
Lyons, W. (1980). *Emotion*. Cambridge: Cambridge University Press.
Oatley, K., and Johnson-Laird, P. N. (1985). 'Sketch for a Cognitive Theory of the Emotions.' *Cognitive Science Research Paper No. CSRP.045*. University of Sussex, Cognitive Studies.
Pylyshyn, Z. W. (1986). *Computation and Cognition: Toward a Foundation For Cognitive Science*. Cambridge, Mass.: MIT Press.
Roseman, I. (1979). 'Cognitive Aspects of Emotions and Emotional Behaviour.' Paper presented to the *87th Annual Convention of the American Psychological Association*.
Ryle, G. (1949). *The Concept of Mind*. London: Hutchinson.
Simon, H. A. (1979). 'Motivational and Emotional Controls of Cognition.' In H. A. Simon, *Models of Thought*, pp. 23–38. New Haven, Conn.: Yale University Press.
Sloman, A. (1969). 'How to Derive "Better" from "Is".' *American Philosophical Quarterly* 6: 43–52.
—— (1978). *The Computer Revolution in Philosophy: Philosophy, Science and Models of Mind*. Hassocks, Sussex: Harvester Press.
—— (1981). 'Skills Learning and Parallelism.' *Proceedings Cognitive Science Conference*, pp. 284–5. Berkeley, Calif.
—— (1984). 'Why We Need Many Knowledge Representation Formalisms.' In M. Bramer (ed.), *Research and Development in Expert Systems*, pp. 163–83. Cambridge: Cambridge University Press.
—— (1985). 'Real-Time Multiple-Motive Expert Systems.' In M. Merry (ed.), *Expert Systems 85*, pp. 213–24. Cambridge: Cambridge University Press.
—— and Croucher, M. (1981). 'Why Robots Will Have Emotions.' *Proc. 7th IJCAI Conference*, pp. 197–202. Vancouver, BC.

11 分布式表述

J·E·欣顿,J·L·麦克莱兰和 D·E·鲁梅哈特*

已知由一些简单的计算元素构成的网络和一些要表述的实体,最直截了当的方案就是将每个实体用一个计算元素来表示。这称为**定位**表述。这样做容易理解,也便于实现,因为物理网络的结构反映出它所包含的知识结构。知识和实现知识的硬件之间的这种关系既自然又简单,因而很多人径直假定,在使用并行硬件时,定位表述是最好方法。当然还存在着许多各种各样的更为复杂的实现方式,在这些实现方式中,不存在概念与硬件单元之间的一一对应,但是,直到这些实现方式产生出那些用定位表述难以取得的效能增长或有价值的自发出现的特性时,它们才受到重视。

本章介绍的表述类型,比起定位表述来,人们不大熟悉,比较难理解。每一实体都由一个分布在许多计算元素上的活动模式来表示,而每一计算元素又与许多不同实体的表述有关。这种比较复杂的表述,其长处并不在于它标记起来很方便,或是容易在常规计算机中实现,而是在于它有效地利用了由简单的、似神经元的计算元素构成的网络的加工能力。

每种表述方案都有它的优点和缺点。分布式表述也不例

外。某些欲得的特性,通过将一些活动模式用作表述方式,会自然而然地产生。另一些特性,像临时存储大量任意联想的能力,实现起来则困难得多。我们将看到,对分布式表述来说,最有力的心理学证据是它们的长处和弱点在一定程度上与人类心灵一致。

本章第一节着重论述分布式表述的某些优点。第二节考虑分布式表述的效能,并清楚地说明了,对一定类别的问题来说,分布式表述之所以优于定位表述的原因。最后一节讨论某些难点,它们是分布式表述的支持者经常回避的问题,诸如成分结构的表述和加工力量在结构目标不同方面的顺序集中。

一些否认声明。在考察支持分布式表述的详尽论据之前，弄清它们在人类信息加工的总理论中所处的地位是很重要的。把分布式表述视作像语义网络或产生系统那种表述方案的**替代物**，是错误的，尽管我们已知这两种表述方式在认知心理学和人工智能中是有用的。把它看作一种以并行网络方式实现那些更抽象的方案的方法，将是更富有成效的，但有一个附带条件：分布式表述将产生某些强有力的、不可预料的自发出现的特性。所以当这些特性在一个较抽象的形式体系中起作用的，可以被看作原素。例如，分布式表述适用

---

* G·E·欣顿等人的"分布式表述"见D·E·鲁梅哈特和J·E·麦克莱兰编辑的《并行分布式处理:认知微结构中的探索》,卷1,《基础》,第77—109页。由麻省理工学院出版社允许重印。本文提到的章节是指《并行分布式处理》一书中的其他章节。

G·E·欣顿(Geoffrey E. Hinton),多伦多大学计算科学系教授。

麦克莱兰(James McClelland),卡内基-梅隆大学心理学教授。

D·E·鲁梅哈特(David E. Rumelhart),斯坦福大学心理学教授。

于按内容寻址的存储、自动概括和最适合当前情况的规则的选择。所以如果我们假定大脑是用分布式表述去实现较抽象的模型，那么把像按内容寻址的存储、自动概括或一个恰当规则的选择这样的能力当作基本运算，不是没有理由的，虽然在常规计算机中没有实现这些运算的简便方法。分布式表述的某些自发出现的特性，在较高层次的形式体系中不容易捕捉到。例如，分布式表述是与同时应用大量部分地适合于当前情况的规则相一致的，而每一规则的应用程度与它的重要性有关。我们将在有关图式的一章(第 14 章)中考察分布式表述的这些特性。在那一章中，我们将清楚地看到，图式和其他较高层次的结构对于依赖于分布式表述的机制只提供了近似的特征描述。因而对分布式表述的分析能够为这些较高层次形式体系作出的贡献是：使某些强有力的基本运算合法化，否则这些运算看起来就像是在演魔术；除了那些在很多较高层次形式体系中能方便地获取到的基本运算之外，进一步丰富了我们所掌握的基本运算的全部内容；指出这些较高层次形式体系可能只获取了基础加工机制的计算能力的粗略特征。

造成混乱的另一个常见原因是这种看法:分布式表述多少是与有关大脑功能定位的广大证据相抵触的(Luria 1973)。为了同时表述完全不同的各种事物,运用分布式表述的系统还需要很多不同的模件。分布式表述存在于这些定位模件的**内部**。例如,不同的模件可供像心理映象和句子结构这些不同的东西使用,但是两种不同的心理映象可能对应于同一模件中的**备选**活动模式。这里提出的表述方式,在全局范围中是局部的,而在局部范围中又是全局的。

1. 分布式表述的优点

这一节考虑分布式表述的三个重要特征:(a)本质上具有构造性的特点;(b)具有自动概括新情况的能力;(c)具有适应变化环境的能力。在这几个优点中,有些也是某些定位模型所具有的,像词汇感知交互激活模型,或第1章中提到的麦克莱兰的概括和恢复模型(McClelland 1981)。

推 理 记 忆

人们有一种十分灵活的存取记忆的方法:他们能从对内容的部分描述中回忆起一些条目(Norman and Bobrow 1979)。而且,即使部分描述的某些地方是错误的,他们也能做到这一点。例如,很多人能迅速回忆起满足下面部分描述的那一条目:一个男演员,聪明人,政治家。这种**按内容寻址的**记忆是十分有用的,但是在常规计算机中很难实现,因为计算机的每一条目是存储在特定地址上的,要取出一个条目,必须知道它的地址。如果用于存取的描述符号组成的所有组合都没有错误,并且是预先已知的,就有可能采用一种称为**傻瓜编码**的方法,在已知一个条目部分内容的情况下,可快速产生这一条目的地址。然而在一般情况下,按内容寻址的记忆在找出最适合这种部分描述的条目时,需要进行大量的搜索。记忆的中心计算问题就是如何使这种搜索有效。当一些线索可能包含错误时,搜索就变得十分困难,因为与这些线索之一

配合失败，就不能用来作快速消除不合适答案的过滤器。

分布式表述提供了一个使用并行硬件去实现最佳配合搜索的有效方法。它的基本思想是相当简单的，虽然它与常规计算机的存储器相差甚远。不同的条目对应于同一组硬件单元上的不同的活动模式。部分描述是用部分活动模式的形式来表现的，这些活动模式激活了某些硬件单元。[①] 于是这些单元之间的相互作用使得这组活动单元去影响其他组的单元，因而就完成了这一模式，并生成最佳配合这一描述的条目。"存储"一个新条目，是靠调整这些硬件单元之间的相互作用，去创造新的稳定的活动模式。它与常规计算机存储器的主要区别在于：不活动的模式不存在于任何地方。它们能够被再创造，因为单元之间的联结强度经过了适当的变化，但是每一联结强度都与存储的多个模式有关，所以不可能指出一个特定位置，说它是用于存储特定条目的存储器。

当弄明白一组简单的加工单元之间的联结能承载大量的不同模式时，许多人都感到惊奇。与分布式模型这个方面有关的说明，参考文献的许多文章里都有(例如 Anderson 1977; Hinton 1981)；在第 17 和 25 章里谈到的记忆和遗忘模型中，也有对这一特性的说明。

有一种考虑分布式记忆的方法，是运用数量很大的一组似然的推理规则。每个活动单元代表一个条目的一个"微特

---

① 如果部分描述不过是一组特征，这是容易的，但是如果部分描述提到与其他对象的关系，就困难得多了。例如，如果要系统追忆约翰的父亲，它就必须表述约翰，但是如果约翰和他的父亲是由同一组单元中的相互排斥的活动模式表述的，就难以明白如何能在不妨碍表述约翰父亲的情况下做到这一点。本文介绍了这个问题的分布式解。

征”,而联结强度则代表微特征之间的似然的“微推理”。这些单元的任何特定的活动模式将满足某些微推理而违反其他微推理。一个稳定的活动模式对似然的微推理的违反程度小于任何邻近模式。通过改变这些推理规则,就可以创造一个新的稳定模式,这样,这一新模式违反这些规则的程度就小于它的相邻模式。有关记忆的这种看法清楚地表明,在真实的记忆与似然的重构之间不存在截然分明的差别。真实记忆是一个稳定模式,因为在它出现之前推理规则就已修改好了。“虚构症”是一个稳定模式,因为推理规则已被修改,可存储数个不同的先前模式。对有关的主体来说,上述情况很可能真假难辨。

真实回忆和虚构症或似然的重构之间的区别有些模糊,这看来就是人类记忆的特征(Bartlett 1932;Neisser 1981)。人类记忆的这种重构性之所以使人感到意外,仅仅因为它是与我们所用的常规比喻相抵触的。我们总是认为,一个记忆系统应该像文件柜或典型的计算机数据库一样,通过存储条目的文字拷贝,然后检索被存储的拷贝这种方式来工作。这种系统不是自然地重构的。

如果我们把记忆看作一种加工,它构造出一个活动模式,该模式代表了与已知线索相一致的最似然的条目,那么我们需要的就是某种保证:它将收敛于与描述具有最佳配合的条目的表述,尽管有时也许容许得到一个好的但不是最佳的配合。想象这种情况是容易的,但是要将它付诸实施就较困难了。解决这一问题的最新方法是使用统计力学去分析多组相互作用的随机单元的行为。这种分析保证了一个条目与这种描述配合得越好,就越有可能作为解答被产生出来。这就是

第7章中介绍的方法,第6章中则介绍了另一种有关的方法。还有一种可供选择的使用连续激活单元的方法(Hopfield 1984),在第14章中介绍。

相似和概括

当一个新的条目存入时,对联结强度的调整不允许清除已存在的条目,这可以通过对大量加权值的稍稍调整来实现。如果这些调整都趋于对已存储的模式有利,那么就会出现一个同向作用的结果:预期模式的总增益将是所有分散的小调整的总和。然而,这对于不相关的模式几乎不起什么作用,因为一部分调整是有益的,而一部分调整是有碍的,它们大都抵消了,所有小调整共同作用的情况是不存在的。这种统计推理成为大多数分布式记忆模型所依托的基础,但是这种基本思想也有很多衍变形式(有几个例子见 Hinton and Anderson 1981)。

将正交活动模式用于各种要存储的条目,有可能完全阻止干扰的产生(第1章中给出了这种情况的基本例子)。然而,这样做就消除了分布式表述的最有趣的特性之一:自动产生概括。如果这一任务仅仅是准确记忆一组无关的条目,那么这种概括效果是有害的,我们称之为干扰。但是在正常情况下,概括是一个有用的现象。它使我们可以有效地处理那些与以前经历过的相似而并不等同的情况。

人们善于对新获得的知识作出概括。如果你获悉有关一个对象的新事实,你对其他类似对象的期望就会发生改变。例如,如果你得知黑猩猩喜欢洋葱,你对大猩猩喜欢洋葱的概

率的估计值就会提高。在使用分布式表述的网络里，这种概括是自动完成的。将关于黑猩猩的新知识包容进来，靠的是调整某些联结强度，使得表述黑猩猩的分布式活动模式的因果作用有所变化。[①] 这种调整自动地改变了所有与之相似的活动模式的因果作用。所以如果大猩猩的表述对同一组单元而言是一种相似活动模式，那么它的因果作用将以相似的方式发生变化。

这个最简分布式方案用的完全是同一组单元上的备选活动模式来表述洋葱概念和黑猩猩概念的。因而难以同时表述黑猩猩和洋葱。要解决这个问题，可以把各别的模件用于一个条目在一个较大结构内的每一个可能的角色。例如，黑猩猩是意中的“执行者”，那么表述黑猩猩的模式就占据了“执行者”模件，而表述洋葱的模式则占据了“承受者”模件（见图11－1）。每个模件都可以具有多个用于所有各种条目的备选模式，所以这种方案与条目的定位表述无关。由定位得到的就是这个角色。

如果你随后又得知长臂猿和猩猩不喜欢洋葱，那么你对大猩猩喜欢洋葱的概率的估计值就会下降，不过也许仍然比最初要高一些。显然，各种事实的组合表明，喜欢洋葱只是黑猩猩的一个特殊的怪癖。一个使用分布式表述的系统会自动地得出这种结论，只要表述各种不同猿猴的备选模式是以特

① 这种模式的内部结构也可能改变。为了使不同输出的联结变得有相关性，是改变输出联结的加权值，还是改变模式本身，在这两者之间始终存在着选择。模式本身的变化改变了它与别的模式的相似性，从而改变了将来怎样进行概括的问题。一般情况下，描绘出如何改变表述一个条目的模式，比描绘出如何改变输出联结，以便一个特定模式在网络的另一部分具有所想望的结果，要困难得多。

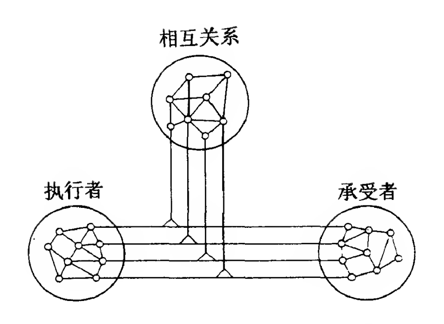


图 11-1 在这个简化图中，有两个不同的模件，其中一个代表执行者，另一个代表承受者。为了体现黑猩猩喜欢洋葱这一事实，一个模件中的黑猩猩模式必须与另一模件中的洋葱模式相联系。“喜欢”以外的相互关系是靠设置第三组单元（其活动模式表述这种关系）来实现的。所以这一模式必然会为执行者组与承受者组之间的相互作用“把关”。欣顿（Hinton 1981）描述了一种使用第四组单元进行这种把关的方法。

定方式相互关联的，这种方式比起仅仅相互相似来显得有点更特殊：每一完整的模式必须有一个部分对所有各种猿猴来说是等同的。换句话说，用于分布式表述的这组单元必须分成两个小组，在第一小组中，所有各种猿猴必然由相同模式表述，而在第二小组中，则由不同模式表述。第一小组上的活动模式表述这一条目的**类型**，而第二小组上的模式表述将区分开属于这一类型的每一例子与其他例子的附加微特征。注意，微特征的任一子集可以看作是定义一个类型的。一个子集可能对所有猿猴来说是共同的，而一个与之不同（但是有交叠处）的子集则可能对所有宠物来说是共同的。因此一个条目可以同时作为许多不同类型的例子。

当这个系统获悉有关黑猩猩的一个新事实时，它通常无法知道究竟这一事实对所有猿猴都成立，或者仅仅是黑猩猩的一种特性。因此显而易见的策略就是调整来自所有活动单元的联结强度，使得新知识部分地成为猿猴的一种普遍特性，部分地成为把黑猩猩与其他猿猴区分开的那些特征的特性。如果接着又获悉其他猿猴不喜欢洋葱，那么就要对调整作出修正，使得有关洋葱的信息不再与所有猿猴的共同子模式相联系。于是这一关于洋葱的知识，将被限制于那个把黑猩猩与其他猿猴区分开的子模式。如果发现长臂猿和猩猩也喜欢洋葱，那么互相得到加强的就是对来自表述猿猴子模式的加权值所作的调整，这种知识就会变得与所有猿猴的共同子模式相联系，而不是与把不同猿猴彼此区别开的模式相联系。

这种概括理论的一个非常简单的形式已在一个计算机模拟中实现(Hinton 1981)。对这种特性所作的几种应用见本书的第Ⅳ部分。

显然可以概括出如下的思想：一个条目的表述是由两部分组成的，一部分表述类型，而另一部分则表述这个特定例子区别于同一类型的其他例子的方式。所有类型本身几乎都是更一般的类型中的例子，通过把表述这一类型的模式分成两个子模式就可以实现这一点，在这两个子模式中，一个代表更一般的类型，上述特定类型只是它的一个例子，另一个代表把这一特定类型与同一个一般类型的其他例子区分开来的特征。因而一个类型与一个例子之间的关系就可以由一组单元和包含着它的一个更大的组之间的关系来实现。注意，类型越一般，用来对它编码的单元组就越小。当内涵描述中的术语数减少时，相应的外延集就增大。

在使用定位表述的传统语义学网络中，概括不是这种表述的直接结果。已知黑猩猩喜欢洋葱，表现这个新知识的显而易见的方法是改变属于黑猩猩单元的联结强度。但这并没有自动地改变属于大猩猩单元的联结。所以为了在定位式方案中实现概括，必须求助于额外的过程。一个通用方法是，使激活从定位单元扩展到表述相似概念的其他单元中去（Collins and Loftus 1975；Quillian 1968）。于是当一个概念单元被激活时，它就会部分地激活它的相邻单元，因此存储于来自这些相邻单元联结中的任何知识就会部分地发挥作用。这一基本思想有许多种不同的形式（Fahlman 1979；Levin 1976；McClelland 1981）。

我们很难明确地将使用带有扩展激活的定位表述的系统与使用分布式表述的系统区分开来。在这两种系统中，激活一个概念的结果都是有许多不同的硬件单元被激活了。在某些模型中，它们的区别几乎完全消失了，比如麦克莱兰（McClelland 1981）的概括模型，其中概念的特性是由特征单元上的一个激活模式来表述的，而这一激活模式是由用作这一概念的例子的潜在地非常多的单元的相互作用所决定的。其主要区别是，在一种情况下，存在着一个独立的特殊硬件单元，它的作用有如“手柄”，使得像概念名称那样的纯常规特性容易接近，同时构造这个网络的理论家也较容易了解网络各部分代表着什么。

如果我们是用人工规定网络中单元之间的联结的方法来构造网络，定位表述方案就有一些明显的优点。首先，比较容易认为，如果我们自己已经为网络提供了所有“知识”，即所有联结，我们就理解网络的行为。但是，如果完成这一工作的是

网络内各个单元之间相互作用影响的整个分布式模式，那么这种理解就常常是虚妄的。其次，人们也许直觉地感到：把任意一个名称加到一个分布式模式上，显然比把它加到一个单个的单元上更加困难。然而，直观上更困难的事，不见得更有效。我们将看到，人们居然能够通过运用分布式表述实现与更少些单元的任意联系。不过在转向这些考虑之前，我们还是先来探讨一下分布式表述的另一个优点：在不配置新硬件的情况下，它们使创造新概念成为可能。

创造新概念

任何用来表述知识的可取的方案都必须具备学习新概念的能力，这些概念在网络开始连线时是无法预先知道的。一个使用定位表述的方案，首先必须独立地决定应在何时形成一个新概念，然后它必须找寻一个备用硬件单元，该单元具有实现这一概念所需的恰当联结。如果我们假定，经过一段时间的早期发展，新的知识是由改变现存联结的强度，而不是由生长出新的联结来实现的，那么也许很难找到这样一个单元。如果每个单元只同其他单元中的一小部分相联结，很可能根本就没有什么单元会同恰好实现新概念的另一些单元相联结。例如，在由一百万个单元构成的集合体中，每一个单元随机地同另外一万个单元相联结，这时，任一单元同包含另外6个单元的特定组相联结的机会只有百万分之一。

为了解脱定位表述的这种困境，几个构思巧妙的方案被提出来了，它们使用的是两类单元。对应于概念的单元不是直接相互联结的，而是由几层中间单元中的间接通道来实现

这些联结(Fahlman 1980; Feldman 1982)。这种方案之所以有效,是因为中间层中潜在通道的数目远远超过物理联结的总数。如果存在 k 层单元,其中每一层对于下一层中随机选择的单元具有 n 个联结分支,那么就存在 n^k 个潜在通道。几乎可以肯定,对任何两个概念单元都将存在一个联结通道,因此沿这一通道的几个中间单元就可以用来联结那两个概念单元。然而,这些方案最后不得不把若干中间单元分派给每个有效联结,而这种分派一旦发生,从每一中间单元发出的多个实际联结中,除了一个以外,其余全都作废了。使用若干中间单元来创造一个单一的有效联结,以此来切换所含元素的单元只有相对较少分支的网络也许是适合的,但是在使用类似大脑的硬件时,这一方法看来就无效了。

找到一个代表新概念的单元,并恰当地给它连线,这种问题在使用分布式表述时是不会发生的。我们要做的只不过是调整单元间的相互作用,以创造一个新的稳定的活动模式。如果是通过对大量联结作非常轻微的调整来做到这一点的,新模式的创造就不一定破坏现存的表述。困难的问题是选择一个适合于新概念的模式。新表述对系统其他部分中的表述产生的作用是由活动单元确定的,因此,关键是使用一批大致具有正确作用的活动单元。通过稍稍改变新模式中活动单元的作用,就可以完成对新模式作用的微调,但为新概念选择一个随机模式,将是不明智的,因为这时还需要对加权值作一些重大的改变,这样就会破坏别的知识。在理想的情况下,为新概念选择的分布式表述应当是这样的:为了使新模式稳定,并使它对其他表述产生符合要求的作用,只需对加权值作出最小的调整。

自然,创造一个新的稳定模式,没有必要一步到位。这一模式的出现可能是在许多各别场合进行调整的结果。这使定位表述中出现的那个棘手问题有所缓和:定位表述系统必须就何时创造一个新概念的问题作出独立的全或无判断。如果我们把概念看作稳定模式,它们的独立性就小得多。例如,通过稍稍调整某些加权值,就有可能将一个稳定模式分化成为两个紧密相关的、但却有所区别的变体。如果不容许我们去克隆硬件单元(以及它们的所有联结),而是用定位表述来完成这种渐进的、概念上的分化,就会困难得多。

分布式表述理论发展中的一个中心问题是,要对学习分布式表述的确切过程作出详细说明。所有这些过程都牵涉到按照第2章概述的那种类型的“学习规则”对联结强度进行的调整。不是所有这些问题都已解决,但是在这些问题上正在取得重要的进展。

2. 行之有效的分布式表述

这一节里,我们考察有关实现分布式表述的一些技术细节。首先,我们指出,某些分布式表述方案可能无法提供区别不同概念的充分根据,并指出,要避免这一局限性需要做些什么。然后,我们介绍了一种方法,可使用分布式表述从联结单元的一个简单网络得到最多的可能存在的信息。其主要结果令人感到惊奇:如果想要用尽可能少的单元对一些特征作出精确编码,使用一些以十分粗略方式调节的单元比较经济,这样,每一特征激活许多不同的单元,而每一单元又由很多不同

的特征激活。于是一个特定特征的编码就是由许多单元中的活动模式、而不是由单个活动单元来完成的,所以粗糙编码是一种分布式表述的形式。

为了使分析简明扼要,我们假定这些单元只具有两种值,开和关。[①] 我们还将略去系统的动态特性,因为从当前来看,我们所关心的问题是,以给定的精度对特征编码,需要多少单元。我们先来考虑这样一种特征:通过给定一个类型(如线段、角、点),以及把这特征与同一类型的其他特征(如位置、方向、大小)区分开来的某些连续参数的值,可以完整地说明这一特征。对于每一特征类型,都存在着一个由可能的例子组成的空间。每个连续参数定义该特征空间的一个维度,而每个特定的特征对应于该空间中的一个点。对于像平面中的点这样的特征,可能特征的空间是二维的。对于像三维空间中有端点、有方向的有界线段这样的特征,特征空间是六维的。我们以考察二维特征空间作为开始,然后推广到更高的维数。

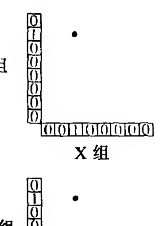
假定我们想要表述平面中单个点的位置,同时又希望用较少的单元达到高精度水平。一个编码方案的精度可定义为当这个点在空间中移动一个标准距离时所生成的不同编码个数。一种编码方案是把这些单元分成 X 组和 Y 组,并指定每一单元对一个特定的 X 或 Y 的区间进行编码,如图 2 所示。这样,一个给定点的编码将从分别来自两个组的两个单元的活动性中得出,而精度是与所用单元的个数成比例的。遗憾

---

① 类似的论据也适用于多值活动水平,但重要的是,不允许多个活动水平具有任意的精度,因为这有可能造成用单一活动水平表述无穷多信息的情况。当一些单元传递一个独立脉冲的概率的变动是单元的活动性的函数时,这些单元看来近似于神经回路中可能存在的那种精确性(见第 20 和 21 章)。

的是，这里存在两个问题。第一，如果两个点必须同时编码，这种方法就无效了。这两个点将激活每一组中的两个单元，而从这些活动单元无法断定这些点究竟是在(x_1，y_1)和(x_2，y_2)，还是在(x_1，y_2)和(x_2，y_1)。这问题称做**结合问题**。它的产生是因为这种表述没有说明什么与什么在一起。

即使每次只有一个点要表述，第二个问题仍然会出现。假定我们想使某些表述而不是别的表述与一个明显的响应相联系：我们要(x_1，y_1)和(x_2，y_2)而不是(x_1，y_2)或(x_2，y_1)与一个响应相联系。我们无法利用从分别代表两个维度的值的单元到响应单元的标准加权联结实现这种联系。因为用于 x_1 的单元和用于 x_2 的单元都必须激活这响应，而用于 y_1 的单元和 y_2 的单元也都必须激活这响应。当用于 x_1 的单元和用于 y_2 的单元都被激活时，是无法阻止这一响应被激活的。这又是因为这种表述无法说明什么必须与什么在一起而造成的结合问题的另一个方面。



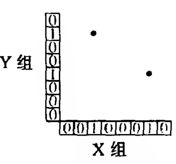


图 11-2 A：使用两组二进制单元对二维空间中一个点的位置编码的简单方法。X 组和 Y 组中的活动单元代表 x 和 y 坐标。B：当两个点必须同时编码时，无法知道哪个 x 坐标与哪个 y 坐标在一起。

在常规计算机中，结合问题是容易解决的。我们只要在计算机的存储器里造两个记录。每个记录包含一对坐标，它们像一个点的坐标一样连在一起，而结合信息则根据两个坐标值处在同一记录中这一事实(通常是指它们处于相邻的存储器位置上)来编码。在并行网络中，

解决结合问题就困难得多了。

联 合 编 码

一种方法是预先为 X 和 Y 值的每一可能组合拨出一个单元。这相当于用大量小的、不交叠的区域覆盖这一平面，每个区域使用一个单元。于是一个点就用单个单元中的活动性来表示，所以这是定位表述。对每一可区分的特征，使用一个单元，由于有一些单元代表两个维度中每一个纬度上的值的联合，结合问题就得到解决。一般情况下，若允许在特定的特征组合与某种输出或其他激活模式之间形成任意的联系，那么都可能需要某种联合表述。

然而，这种定位编码的代价很大。比起前面的方案来，它的效率要低得多，因为对平面中一个点的准确描述的精度只与单元数的平方根成正比。一般情况下，对一个 k 维特征空间来说，这种定位编码产生的精度与单元数的 k 次根成正比。在不涉及结合问题的情况下，这样得到高精度的代价是很大的。

如果在每一情况中，都有数量非常之大的特征出现，那么对每一个可区分特征使用一个单元，也许是一种合理的编码，这样，这些单元之中有一大部分是活动的。但是，如果同时出现的只是可能特征之中一个很小的部分，这种编码的效率就非常之低。如果这种单元只在一半时间里是活动的，由一个二进制单元的状态传递的平均信息量就是一个比特，而如果这个单元只偶尔是活动的，平均信息量就小得多了。① 因此，

① 开通的概念是 p 的一个单元传递的信息量，是$-p\log p-(1-p)\log(1-p)$。

使用一种在任何时刻活动单元都占较大部分的编码，将会是较有效的。如果我们放弃由单个单元的活动性来表述每个可区分特征的思想，这一点就能够实现。

粗糙编码

假定我们把空间分成较大的、交叠的区域，并且给每个区域分配一个单元。为简单起见，我们假定这些区域是圆形的，它们的圆心均匀地随机分布于整个空间，并且给定的编码方案所使用的所有区域都具有相同的半径。我们关心的问题是，一个特征以何种精度被编码为这些区域的半径的函数。如果有给定数目的单元供我们使用，那么是用大区域，使得每个特征点落入多个区域里较好呢，还是用小区域，使得每个特征由个数较少的、但调节较细微的单元中的活动性来表述较好呢？

精度是与当我们使一个特征点从空间的这一边到另一边沿直线移动时生成的不同编码的个数成正比的。这条直线与区域边界每交叉一次，特征点的编码就发生变化，因为对应于这一区域的单元的活动性改变了。所以沿这条直线的可区分特征的个数，刚好两倍于这条线穿过的区域个数。[①] 这条线穿过的每一个区域的圆心都位于线两旁的半径范围以内（见图 11-3）。这个数目是与区域半径 r 成正比的，它也与区域

① 如果你进入并离开一个区域，这期间没有穿越其他区域的边界的话，会出现一些问题，因为你回复到了以前的同一编码，但是如果这些区域足够密集，因而存在着许多区域，包含了空间中每一个点，那么这种影响就可忽略。

的数目 n 成正比。因而精度 a 与区域数和区域半径的关系如下：$a \propto nr$。

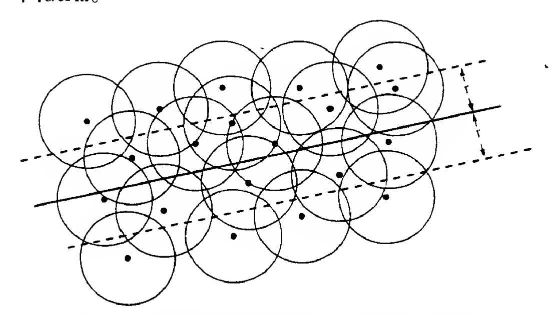


图 11-3 被直线切割的区域边界数与直线周围区域半径范围内的区域圆心数成正比。

一般来说，在 k 维空间中，其中心处于通过空间的一条直线的半径范围以内的区域的个数是与半径为 r 的一个 k 维超柱体的体积成正比的。该体积等于这个柱体的长度（是确定的）乘以它的（k-1）维横断面积，该面积与 r^{k-1} 成正比。因此，精度由下式给出：$a \propto nr^{k-1}$。

这样，例如当区域半径增加一倍时，表述一个像有端点、有方向的三维棱那样的六维特征的**线性**精度就会增大到 32 倍。较大区域导致较粗率的表述的直觉看法是完全错误的，因为分布式表述对信息的掌握比定位表述要有效得多。即使每个活动单元的意义欠明确，活动单元的组合的意义却明确得多。还需注意的是，使用粗糙编码时，精度是与单元数成正比的，这比与这数的 k 次根成正比好得多了。

在视觉脑皮层中，响应视网膜映像的复杂特征的单元，常

常有着相当大的接受场。这常被解释为走向平移不变表述方式的第一步。然而，这些大型场的功能也可能不是实现平移不变性，而是精确指出这个特征在哪里！

粗糙编码的局限性。至此，仅仅谈到粗糙编码的优点，而忽略了它有问题的一些方面。当"接受场"变得太大时，就会出现许多限制，以致造成粗糙编码策略的失败。当这种场在大小上变得相当于整个空间时，会出现一个明显的限制。这一限制一般不起什么作用，因为在接受场变得这样大之前，别的更严峻的问题就出现了。

只有在必须表述的特征相对稀疏的时候，粗糙编码才是有效的。如果许多特征点拥挤在一起，每个接受场将包含许多特征，同时这些粗糙编码单元中的活动模式将不能在特征点的许多备选组合之间作出区分。(如果使这些单元具备反映落入它们接受场内的特征点的个数的整体活动水平，那么只有附近少数几个点可以被接受，而不是许多点。)因而存在着分辨率/精度的权衡比较。如果特征分得很开，因而高分辨率也并非需要，那么粗糙编码就能给特征参数以高精度。作为一种粗浅的经验，接受场的直径应当与同时出现的特征点之间的间隔具有相同的数量级。①

如果已知粗糙编码之所以优于定位编码，是因为它通过使每个单元更经常地活动，从而更有效地利用了单元的信息携带能力，那么对粗糙编码只有在特征点稀疏的情况下才是有效的这一事实，就不应该感到奇怪了。如果特征点十分密

---

① 有趣的是，许多几何学上的视错觉印证了在相隔距离比主体特征位置认识中的不确定性大得多的情况下的特征之间的相互作用。这正是使用粗糙编码精确表述复杂特征时可望出现的情况。

集,以致这些单元在使用定位编码的大约一半时间里是活动的,粗糙编码只能使事情更糟。

使用粗糙编码的第二个主要限制源于这一事实:一个特征的表述必须用于对其他表述产生影响。如果在这些特征能够对其他表述产生恰当作用之前,必须将它们重编码为微调单元中的活动性,那么使用粗糙编码就失去了意义。如果我们假定一个分布式表述的作用是组成这一表述的各个活动单元的作用之和,那么对于可以有效使用粗糙编码的条件就存在着一个很强的限制。相近的特征将由相似的活动单元组来编码,因此它们将不可避免地趋于有相似的作用。大体上说,只有当一个特征所需的作用是相邻特征所需作用的平均值时,粗糙编码才是有用的。在足够细小的尺度上,对于空间作业来说,这几乎总是成立的。使它失效的那一尺度确定了接受场大小的上限。

另一限制是,每当粗糙编码的表述方式相互作用时,就存在着粗糙性增长的趋势。要减少这种趋势,可能需要在每个表述内部进行横向抑制运作。这个问题需要进一步研究。

向非连续空间的扩展。通过把一组条目看作一个接受场的等价物,作为粗糙编码基础的原理就可以推广到非连续空间。定位表述对每一可能存在的条目使用一个单元。分布式表述将一个单元用于一组条目,并以隐含方式将一个特定条目编码为几个活动单元对应组的相交。

在空间特征的领域中,规律性一般是很强的:有着相似参数值的特征的集合,对其他表述的作用必须是相似的。粗糙编码之所以有效,是因为它允许这种规律性表现在联结强度中。在其他领域里,规律是不同的,但有效性的论据是相同

的：把一个单元用于一组条目比用于单一条目更为有利，只要这组条目是以如下方式选择的：该组的成员资格与其他组的成员资格有所牵连。于是，就可以将这种牵连关系作为联结强度。理想的情况是，组的选择应当使本组的成员资格与也是由个体单元编码的另外一些组的成员资格具有强牵连关系。

我们用一个十分简单的例子来说明这些论点。我们考察一种微型语言，它是由三个字母的英语单词组成的，其构词方式是在 w 或 l 后面接 i 或 e，再在后面接 g 或 r。字母串 wig 和 leg 是单词，而 weg，lig 及所有以 r 结尾的字母串都不是。假定我们想要使用一个分布式表述方案作为表述这些单词的基础，并且我们希望能够使用分布式模式作为判定一个字母串是单词还是非单词的基础。为简单起见，我们将使用单一的"判断"单元。问题是找出从表述单词的单元到判断单元的联结，使得只要一个单词出现时，判断单元就启动，而出现的不是单词时，判断单元就不启动。①

图 11－4 表示三种表述方案：一个不起作用的分布式方案、一个起作用的分布式方案和一个定位式方案。在第一种方案里，每个字母/位置组合是由不同单元代表的。由于只有五种字母/位置组合的可能性，所以只有五个单元具有与输出单元的联结。每个单词和非单词在这五个单元之上产生一个

① 注意，如果这个判断单元由一组单元替代，并且该网络的任务是产生一个关于单词和非单词判断的不同模式，这个问题仍然如此。因为当我们考查每个单元时，该单元在这两种模式中或是取相同的值，或是取不同的值；在值相同的情况下没有问题，但是这样的单元同时也不区分这两种模式。当值不同时，单元的行为正如本文讨论过的那种单一判断单元一样。

不同的并且是唯一的模式，但是从这五个单元到判断单元的联结，不能以这种方式建立：只要有一个单词出现，判断单元就启动，同时只要有一个非单词出现，判断单元就不启动。

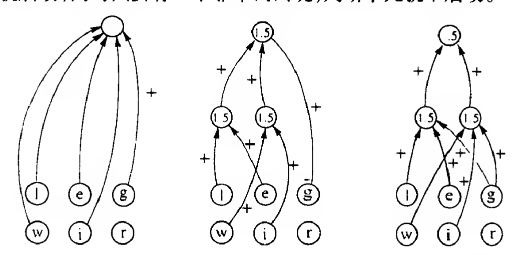


图 11-4 确定由字母 w 或 l 后接 i 或 e 再接 g 或 r 而构成的字母串中哪一些形成单词的问题所用的三种网络。联结上的数字表示联结强度；单元上的数字表示这些单元的阈值。如果单元的输入超过它的阈值，单元就取得等于 1 的活动性；否则它的活动性是 0。

产生这个问题的原因很简单：字母/位置单元与判断单元之间的联结所能达到的程度，仅限于每个字母指明这个字母串是不是一个单词的程度。g 很可能表明一个单词的出现，而 r 则表明这个条目不是单词；但是在这个例子中，如果分别地看待其他每一个字母，它们都绝对不具有预言能力。

一个字母串是不是一个单词，是不能从字母串所包含的各个字母中得出明确结论的；还必须考虑它包含哪些字母组合。因而，我们需要一种表述，它以充分满足网络目标的方式获得字母组合所表现的东西。我们可以通过使用定位表述，以及为每个词分配一个节点的方式，像图 11-4 中的第 3 图那样，来获得这一点。然而，重要的是要看到：我们没有必要

始终采用定位表述方法来解决我们网络的问题。联合分布式表述是能做到这一点的。

图中第2图表示的方案提供一个联合分布式表述。在这方案中,有一些用于成对字母的单元,这些单元在这有限的词汇中碰巧获得了对于确定一个字母串是不是一个单词来说至关重要的那些组合。当然,它们是 wi 对和 le 对。这些联合单元,连同从 g 单元到判断单元的直接输入,足以构成一个网络,它对所有由 w 或 l 后接 i 或 e 再接 g 或 r 所组成的字母串作出正确的分类。

这个例子说明,如果要用分布式表述来解决那些很可能对网络提出的问题,联合编码常常是必不可少的。这一论点可以用许多别的例子来说明——**不可兼或**问题就是个经典的例子(Minsky and Papert 1969)。需要某种联合编码的问题的其他一些例子见欣顿著作(Hinton 1981)以及第7,8两章。联合编码对心理学模型的应用见第18章。

在根本不用任何联合编码的情况下,某些问题(大多是非常简单的问题)也能得到解决,而另一些问题将需要每次多于两个单元的联合。一般很难事先准确规定所需联合的"阶"是什么。较好的办法倒是找到能发现恰当表述的一种学习方案。第7和第8章中提出的机理,是朝着这个目标迈出的两步。

实现两个领域之间的任意映射

留心的读者一定注意到了:在我们刚刚考察过的例子中,采用**定位**表述总是有效的。然而,我们已经探讨了分布式表

述所以更为可取的几个理由。其一是它们能比定位表述更有效地使用并行硬件。

这一节要说明的是,一组单元中的分布式表述怎样能引起另一组单元中的一个恰当的分布式表述。我们来考察在这两个组中实现表述间的任意配对的问题,我们采用的例子是前面例子的扩展:一个单词的直观形式与它的意义之间的联系。所以要考察一个任意映射,是因为这是定位表述看起来最有用的场合。如果在这场合分布式表述更好,那么在一些存在着基本规律性,而这些规律可以通过一组单元和另一组单元上的活动模式的规律性来获取的场合,分布式表述肯定更好。关于在这些场合分布式表述所具有的优越性的讨论,见第 18 章。

如果我们把问题限制于单语素结构的单词,那么从知道何种字母串意指什么无助于我们预测一个新字母串意指什么的意义上,从字母串到意义的映射具有任意性的特点。[①] 从字母到意义的映射中的这种任意性,把似然性赋予了具有显式词单元的模型。显然,如果存在这样的单元,任意映射就能实现。一个字母串正好激活一个词单元,而这又激活了我们希望与之联系的任何一种意义(见图 11 - 5A)。于是相似的字母串的语义可能是完全独立的,因为它们是由分离的词单元作中介的。这里完全不存在作为分布式表述特性的自动概括。

---

① 甚至对单语素结构的单词来说,也可能存在与意义有联系的特定片断。例如,以 sn 开头的单词,通常指以嘴唇或鼻子所表现出的使人不愉快的样子(讥笑、咆哮、窃笑),而带长元音的单词比带短元单的单词更有可能表示大的、慢的东西(G. Lakoff,个人通信)。L·卡罗尔(Lewis Carroll)的许多诗靠的就是这些效果。

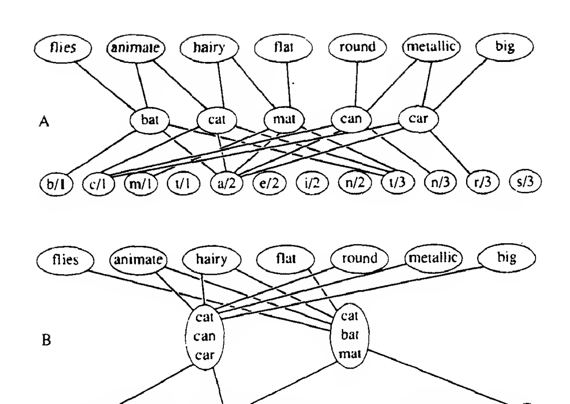


图 11－5　A：一个三层网络。底层包含的单元代表单词内部特定位置上的特定字母。中层包含识别完全单词的单元，而顶层包含的单元代表单词意义的语义特征。这个网络在中层使用单词的定位表述。**B：**顶层和底层与(A)相同，但是中层使用一个更大范围的分布式表述。该层的每个单元都能由一个完整的单词集合中任何一个单词的字母表述来激活。因而，这个单元为在激活它的那些单词中的任何一个的意义中出现的每一个语义特征提供输入。本例中只示出了那些包含单词 cat 的单词集合。注意，接收来自所有这些单词集合的输入的语义特征只是 cat 的语义特征而已。

如果在一个系统中，单元的中间层把单词编码为分布式活动模式，而不是编码为单个定位单元中的活动，那么从直觉上就完全看不出这种系统能实现任意映射。分布式备选方案看来有严重的缺点。一个活动模式对其他表述的作用是这一模式中许多活动单元个别作用合起来的结果。所以相似的模式往往有相似的作用。看来我们不能随意地使一个给定的模

式具有我们所希望的对意义表述的作用,而不同时改变其他模式所具有的作用。看来这种相互作用使得从单词的分布式表述到意义表述的任意映射难以实现。现在我们要说明的是,这些直觉是错误的,单词的分布式表述可以非常有效地工作,甚至可能比单个词单元更为有效。

图 11 - 5B 显示一个三层系统,其中字母/位置单元馈入**单词集合**单元,接着该集合单元又馈入**语义**或**义素**单元。这一类型的模式以及与之关系密切的变体,已经由威尔肖(Willshaw 1981)、V·多布森(V. Dobson, 个人通信 1984)和 D·齐普泽(D. Zipser,个人通信 1981)分析过;某些与之有关的进一步的分析在第 12 章中讨论。为简单起见,我们假定每个单元或是活动的,或是不活动的,并且不存在反馈或交叉联结。这些假定也可以放宽,而不会对论证产生实质性的作用。每当字母/位置单元的模式为一个特定集合中的一个单词编码时,一个单词集合单元就被激活。例如,这集合可能是所有以 HE 开头的四字母单词,或是所有至少包含两个 T 的单词。这里所要求的只不过是,能够通过应用对被激活的字母/位置单元的简单测试,来确定一个单词是否在这一集合中。所以例如所有意指"美好"的单词的集合,就不可能成为一个单词集合。这里有一个隐含的假定:词意要能够表示为义素集合。这一论点尚有争议。成分分析的观点认为,意义是一组特征,结构主义的观点认为,单词的意义只能根据它与其他意义的**关系**来定义,这两种观点之间看来存在着一道鸿沟。我们在本章稍后的地方考察一种把这两种观点结合起来的方法,其做法是让接合的表述方式由活动特征的许多不同的集合构建而成。

回到图 11 - 5B,这里的问题是,当词集单元分别由多于一个的单词激活时,是否可能实现字母/位置矢量与义素矢量之间的任意一组联系。在许多可能的特定模型中,仅考察一个就够了。让我们假定,一个活动的词集单元给所有存在于单词集合的任何一个单词的意义中的义素单元提供正输入。让我们再假定每一义素单元具有一个可变的阈值,它被动态地调整到比活动词集单元的个数恰好稍微小一点。这样,只有那些正在接收来自每一活动词集单元的输入的义素单元,才会成为活动的。

正确单词的所有义素都会被激活,因为每一个这种义素都将出现在活动单词集合的一个单词的意义中。然而,附加义素也可能被激活,因为它们可能完全是碰巧地接收到来自每一个活动词集单元的输入。要使一个义素接收的输入小于它的阈值,至少必须有一个活动单词集合不包含任何以这种义素作为其部分意义的单词。对每个活动单词集合来说,发生这种情况的概率 i 为 $i=(1-p)^{(w-1)}$,其中 p 是包含这一义素的单词的比例,而 w 是词集单元的单词集合中单词的个数。以 w - 1 为幂的原因是已经假定这义素不作为正确单词的部分意义,所以只剩下 w - 1 个单词能将这义素保留在它们的意义中。

假定当一个单词在字母层次上编码时,它激活了 u 个词集层次上的单元。每一个不作为单词部分意义的义素,不能从每个词集单元接收输入的概率是 i。因此,所有这些词集单元将给它提供输入的概率 f 就是

$$f=(1-i)^{u}=[1-(1-p)^{(w-1)}]^{u}。$$

据检查,当 w 等于 1 时,一个“错误肯定”的义素的这一概

率减小到零。表 11-1 显示在 p,u 和 w 值的各种组合下的 f 值。注意,如果 p 非常小,即使 w 相当大,f 仍可以忽略不计。这意味着,如果语义特征是相对稀疏的,即每一词义仅包含整个义素集合中的一小部分,那么每个词集单元都参与许多单词的表述的那种分布式表述就不会导致错误。所以在义素单元是全然独特的(不为过多的不同词义所共有的)情况下,词集单元可以是全然非独特的。表中的某些项目清楚地表明:对某些 p 值,即使词集单元个数大大小于单词数(单词与词集

表 11-1

u	w	p	f	u	w	p	f	u	w	p	f
5	5	0.2	0.071	5	5	0.1	0.0048	5	5	0.01	9.5×10^{-8}
5	10	0.2	0.49	5	10	0.1	0.086	5	10	0.01	4.8×10^{-6}
5	20	0.2	0.93	5	20	0.1	0.48	5	20	0.01	0.00016
5	40	0.2	1.0	5	40	0.1	0.92	5	40	0.01	0.0036
5	80	0.2	1.0	5	80	0.1	1.0	5	80	0.01	0.049
10	10	0.2	0.24	10	10	0.1	0.0074	10	10	0.01	2.3×10^{-11}
10	20	0.2	0.86	10	20	0.1	0.23	10	20	0.01	2.3×10^{-8}
10	40	0.2	1.0	10	40	0.1	0.85	10	40	0.01	1.3×10^{-5}
10	80	0.2	1.0	10	80	0.1	1.0	10	80	0.01	0.0024
10	160	0.2	1.0	10	160	0.1	1.0	10	160	0.01	0.10
40	40	0.2	0.99	40	40	0.1	0.52	40	40	0.01	2.7×10^{-20}
40	80	0.2	1.0	40	80	0.1	0.99	40	80	0.01	3.5×10^{-11}
40	160	0.2	1.0	40	160	0.1	1.0	40	160	0.01	0.00012
40	320	0.2	1.0	40	320	0.1	1.0	40	320	0.01	0.19
40	640	0.2	1.0	40	640	0.1	1.0	40	640	0.01	0.94
100	100	0.2	1.0	10	100	0.1	0.99	100	100	0.01	9.0×10^{-21}
100	200	0.2	1.0	10	200	0.1	1.0	100	200	0.01	4.8×10^{-7}
100	400	0.2	1.0	100	400	0.1	1.0	100	400	0.01	0.16
100	800	0.2	1.0	100	800	0.1	1.0	100	800	0.01	0.97

错误肯定义素的概率 f,是按每个单词计的活动词集单元数 u、每个单词集合中的单词数 w 和义素作为部分词义的概率 p 的函数。

单元之比是 w/u)，出错的机会也是可以忽略的。

上述例子作了许多简化假定。例如假定每个词集单元是与每个相关义素单元相联结的。如果失去任何一个这样的联结，我们就无法为这些义素单元提供与活动词集单元数相等的阈值。考虑到失去联结的情况，我们可以降低阈值。这将提高错误划类的出错率，但是这作用也许很小，并且通过增加词集单元，使得单词层次表述的独特性有所提高，这作用可得到补偿(Willshaw 1981)。或者我们可以使每个词集单元否决那些在单元的任何单词中都不出现的义素。这一方案有效地抵制了联结的消失，因为如果存在着其他否决，缺少一种否决是可以允许的(V. Dobson，个人通信 1984)。

还有两种简化假定，就任意映射作业而言，这两者都造成对分布式表述有效性的低估。首先，这些演算假定，在最经常出现错误的场合，根本不存在用增加一些加权值并减少另一些加权值来改善性能的微调过程。第二，这些演算略去了存在于义素之间的交叉联结。如果每一词义都是一些义素的一个常见的稳定模式，那么就存在着一个强"清理"作用，对一个特定的词义来说，一旦义素层次上的激活模式充分接近常见模式，该作用势必抑制错误的义素。义素之间的相互作用也为单一字母串(例如 bank)引出两种完全不同意义的能力提供了解释。激活词集单元**自下向上**的作用对两种义素集合都有帮助，但是在**自上向下**的因素对一种意义有利的同时，另一意义中的义素将被义素层次上的竞争性的相互作用所抑制(Kawamoto and Anderson 1984)。

模拟。只要义素单元中间存在交叉联结，并且存在为了避免频繁出错而对各个加权值进行的微调，上面给出的较为

简明的概率分析就会失效。为了给交叉联结留出时间去清理输出，必须使用一个迭代过程，而不是简单的“直通”加工，在这种加工中，每一层都用单一的同步步子完全确定下一层中的所有单元的状态。含有交叉联结、反馈和非同步加工元素的系统很可能是更符合实际的，但是一般也是很难分析的。然而，我们现在开始发现，存在着这些较复杂系统的子类，它们的行为方式是易于处理的。第 7 章中较详细地叙述了一个有关这种子类的例子。它使用了一些表现出固有随机特性的加工元素。出乎意料的是，采用随机元素之后，这些网络得以**更好**地执行搜索，**更好**地学习，也**更容易**分析。

这种简单网络可用来说明某些要求，一种要求是关于运用义素单元之间的相互作用来“清理”输出的能力的，另一种是关于通过对合适的加权值作微调来避免错误的能力的。这种网络含有 30 个字母单元、20 个词集单元和 30 个义素单元。字母与义素单元之间不存在直接联结，但每个词集单元是与所有字母和义素单元相联结的。字母单元按每组 10 个分成 3 组，同时每个 3 字母单词在每一 10 字母组中有一个活动单元（单元的活动水平只能是 1 或 0）。选定每个义素单元活动性的概率为 0.2，就可以随机地挑选一个单词的“意义”。图 6 所示网络已经学会将 20 个不同的字母串连同选定的意义联系起来。每个词集单元都与许多单词的表述有关，而每个单词又牵涉到许多词集单元。

用来创造这个网络的学习过程和当给出字母输入时用来定出一组活动义素的搜索过程，第 7 章中有详细说明。这里，我们对这种模拟的主要结果作一简单总结。

经过一段较长时期的学习，这种网络在给出一个字素输

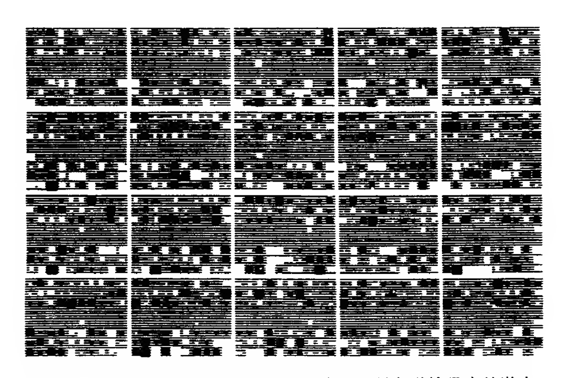


图 11-6 位于三层网络中层的 20 个单元的所有联结强度的激光显示。这个网络能把底层(表示字母)的 30 个单元上的活动模式映射到顶层(表示义素)的 30 个单元上的相联系的活动模式上。在用来描述中层单元的每个大矩形里,顶部的 30 个黑白矩形描述与顶层联结的加权值,而底部的 30 个矩形描述来自底层的加权值。白矩形是正加权值,黑矩形是负加权值,巨型面积表示加权值的大小。出现在单元中部某处的单个加权值是它的阈值(黑色指正阈值)。顶层 30 个单元之间的加权值没有在图中标明。

入时 99.9%地产生正确的义素模式。经学习后,移走词集单元中的任何一个单元,常常会引起几个不同单词的出错率的轻微升高,而不是完全失去一个单词。在别的分布式模型中也已观察到类似的结果(Wood 1978)。在我们的模拟中,有些错误响应是相当有趣的。在对丢失一个词集单元的模型所做的 10000 次测试中,该模型未能恢复正确义素模式的例子有 140 个。其中有些是由失去的或额外的 1 或 2 个义素组成的,但是在出错情况中有 83 个恰好是某一别的单词的义素模式。这是义素单元之间相互协调作用的结果。如果来自

词集单元的输入是噪声或不明因素，即与单元失去作用时的情况一样，那么清理的结果有可能定出一个相似的、但不正确的意义。

这一结果使人想起被称为**严重诵读障碍**的现象，它是成年人因为某种脑损伤而发生的。在出示一个单词并要求读出它时，受试者有时候会说一个意义非常近似的不同单词。这不正确的单词有时有着十分不同的发音和拼法。例如，当出示单词 PEACH[①]时，受试者会说成 APRICOT[①]。(有关后天诵读障碍的进一步的资料，见 Coltheart, Patterson and Marshall 1980。)这类语义错误似乎是很奇怪的，因为看起来受试者在造成语义联系的错误时，肯定已经接触到了词条 PEACH，可是如果他能够碰到这个词条，为什么不能说出它呢？(这些受试者也许知道并有能力说出他们读错的那些词。)采用分布式表述，我们就不必严格区分是否接触到一个单词了。在一个学习过单词 PEACH 的网络中，PEACH 的字母表述将近似地使这些义素单元得到正确输入，所以义素层次上的相互作用恰好产生 APRICOT 的义素模式。另一个心理学上有趣的结果出现在网络破坏后又重新学习的时候。对每一个涉及词集单元的联结增加噪声，这个网络就会受到破坏。这样就使性能从 99.3%正确降低到 64.3%。[②] 然后重新对这个网络进行训练，它就会出现十分快的重新学习过程，比它在性能是 64.3%时原来的学习效率要快得多。几何学论据对这种快速

① PEACH 意为桃子，APRICOT 意为杏子。——译者

② 在这例子中，出错率(似应为“正确率”——译者)是 99.3%，而不是 99.9%，因为这种网络必须较快地作出响应，所以共同的作用使定出最佳输出所用的时间比较少。

恢复作出了预测,该论据表明,存在着有关一组联结强度的某种特殊的东西,这是通过向一组近似完备的联结强度增加噪声而生成的。最后得到的这组联结强度与表现出同样性能的其他组的联结强度有很大的不同。(进一步的讨论见第 7 章。)

如果在再训练时将几个单词删去，会有更意想不到的结果出现。在进行再训练的过程中，这些单词的出错率大大地减少了，尽管其他字母-义素配对与它们并没有内在关系，因为所有这些配对都是随机选择的。网络没有再次显示单词的“自发”恢复，这是使用分布式表述的结果。所有这些加权值都与再训练中显示出的那些单词的子集合的编码有关，因此就会从每个加权值中消去这种后增的噪声。对每个单词使用分离单元的方案不会有这种表现，所以我们可以把未训练过的条目的自发恢复看作分布式表述的一个性质表征。

3. 结构表述和加工

这一节里,我们来看看分布式表述的两种扩展。这些扩展说明,分布式表述的思想与人工智能领域有关结构在表述和加工中的重要性的某些主要见解是一致的。也许因为分布式表述的某些倡导者没有特意与这些论点相协调,所以常常看不清楚在分布式表述方案中结构是怎样获取的。本节的这两部分指出了在扩展分布式表述以处理这些重要问题时可以采取的某些方向。

成分结构的表述

任何系统如果试图实现人类使用的各种概念结构，都必须具备表述两种颇为不同的层级体系的能力。第一种是“IS－A”体系，它使类型与它们的例子相联系。第二种是部分/整体体系，它使条目与组成它们的成分条目相联系。IS－A体系的最重要的特征是类型的已知特性必须被例子“继承”，而已知可应用于一个类型的所有例子的那些特性，必须正规地从这类型得出。在本章的前面部分，我们已看到，怎样可以通过使一个例子的分布式表述**包含**这一类型的分布式表述（作为子部分），来实现IS－A体系。这种表述技巧自动产生出IS－A体系的最重要的特征，但是这种技巧只能用于一种层级体系。如果我们用活动模式之间的部分/整体关系来表示条目之间的类型/例子关系，看来就不能也用它来表示条目之间的部分/整体关系。我们不能使整体表述变成它的部分表述的总和。

如何表示一个条目与组成它的成分条目之间的关系的问题，已成为那些假定分布式表述成立的理论的一个主要绊脚石。在与之对立的定位式方案中，一个整体就是一个节点，该节点由标示弧与表示它的部分的节点相连接。但是分布式方案的中心宗旨是，不同条目与同一组单元中的各种**备选**活动模式相对应，所以看来整体和它的部分似乎是不能同时表述的。

欣顿（Hinton 1981）提出了一个摆脱这一困境的方法。该方法基于整体不是各部分简单加和这一事实。整体是由那些在整个结构内部扮演特殊角色的部分构成的。例如，一个形

状是由一些具有特定尺寸、方向和位置的相对于整体来说较小的形状构成的。每一个成分形状都作为一定的空间角色，而整个形状是由一组形状/角色对构成的。[①] 类似地，一个命题是由在整个命题结构中具有特定语义角色的一些事物构成的。这就提供了一种实现整体与部分之间的关系的方法：每个部分的本体(identity)首先应该与它的角色组合，产生出表示本体与角色组合的单一模式，然后整体的分布式表述应由这些本体/角色组合的分布式表述(以及某些附加的"自发"特征)之和来组成。这一提法不同于那种简单地认为整体的表述是它的部分的表述之和的思想，因为用来表示本体/角色组合的子模式与用来单独表示这些本体的模式有很大区别。例如，它们并不把这些模式作为部分包含进去。

自然，一个项目作为一个整体按它自己的实际情况的表述，与在较大结构内部扮演特定角色的同一项目的表述之间，必定存在着一条通路。例如，必须有可能从两个分离的、显式的分布模式形成本体/角色的表述，其中一个模式表示本体，另一个表示角色。必须也有可能找到另一种方式，从本体/角色组合的单个组合表述生成本体和角色的显式表述(见图11－7)。

使用本体/角色组合的表述模式，使部分/整体体系可以按同样的方式表示为类型/例子体系。我们可以把这整体简单地看作许多更一般的类型中的一个特例，每一个这种较一

① 部分之间的关系也是很重要的。将形状/角色对清楚地表示出来的一个好处是，它使不同的对可以互相支持。人们可以把一个物体内部各种不同的位置看作槽，而把一个物体的形状或部分看作这些槽的填充物。这样，对整个形状的知识就可以由各种不同的填充物之间的正相互作用来实现。

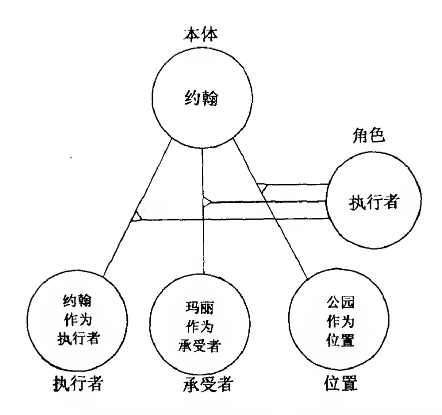


图 11－7 把本体和角色的分离表述组合成为单一模式时可能需要的组织的示意图。每次只能显式表示一个本体和一个角色，因为每一本体组和角色组一次只能有一个活动模式。然而，各种不同的角色组允许多个本体/角色组合同时被编码。图中小三角符号表示活动模式的一种能力，该活动模式在那个显式表述角色的组中，而这种能力则在众多角色组中确定出哪一组当时正在与本体组相互作用。这样就使具有特定角色的本体可以被"读出"，同时也允许将本体与角色组合起来的反向运作存在。

般的类型都可以定义为具有一个扮演特定角色的特种部分的类型（例如装有木制腿的人）。

顺序的符号加工

如果成分结构是以上述方式实现的，就存在一个严峻的问题：在任一时刻能有多少结构是活动的。配置硬件的一个显而易见的方法是，对一个结构内部每一可能的角色使用

一组单元，并且使这个组中的活动模式表示当前正在扮演这角色的那个成分的本体。这意味着，除非我们愿意为这一整体结构假定多重拷贝，否则每次只能表示一个结构。使用一些带有编程联结而不是固定联结的单元来做到这一点的一种方法在第16章中介绍。但是，如果有数量较多的模件必须同时"编程"的话，即使这种技术也会碰到困难。然而人们看来的确在能够同时处理的一般相同类型的结构个数方面受到很强的限制。在已知大脑具有大规模并行结构的情况下，人们在这种高描述层次上表现出的顺序性，一开始是使人感到惊讶的，但是如果我们放弃定位式的先入之见，代之以分布式方案，对这一现象的理解就变得容易多了，这种方案采用的是并行关系，它给予每一活动表述一个非常丰富的内部结构，这些结构使得正确种类的概括和按内容寻址成为可能。如果认为每个"符号表述"等同于一个大的互作用网络的连续状态，那么人类是顺序**符号**加工者的看法就有几分道理。对这些问题的进一步的讨论，见第14章。

顺序符号加工方法的一个中心宗旨(Newell 1980)就是要具备这样的能力：即能够集中于结构的任何部分，并把它扩展为一个整体，该整体与以它作为一部分的那个原来的整体在内容上同样丰富。把一个结构的各部分扩展到数目不限的层次上去的递归能力，以及把整体结构压缩成简化形式，使整体结构可以用作更大结构的组成成分的反演能力，是符号加工的本质。有了这种能力，一个系统可以从那些代表另一些整体结构的东西中建造出许多结构，而不要求将这些另外的结构的繁琐的细节全部表示出来。

在常规的计算机执行过程中，这一能力是通过使用指示

器来获得的。指示器很方便，但是它们依赖于使用地址。在并行网络中，为了实现符号加工，我们需要在功能上等价于任意点的某种东西。这恰好是代表本体/角色组合的子模式所提供的。这些子模式使某个部分的整个本体可以从整体的表述和系统希望集中注意的那个角色的表述中取得，同时它们也使得一个本体和一个角色的显式表述可以被组合成一个不大繁琐的表述，这样，数个本体/角色组合就可以同时被表述，以便构成一个较大结构的表述。

总　　结

给定一个并行网络，条目的表述可以通过单一定位单元的活动，或通过一个大的单元组中的活动模式，在这单元组中，每个单元都为条目的一个微特征编码。只要存在着一些可以通过微特征之间的相互作用获取的基本规律性，分布式表述就是有效的。通过将每一小段知识编码为一个大的相互作用集合，就有可能获得像按内容寻址的记忆和自动概括那样的有用特性，并且能创造新的条目，而不必创造硬件层次上的新的联结。在连续变动的空间特征的领域中，为分布式表述使用中的优缺点提供一个数学分析是相对容易的。

分布式表述似乎不适合于实现纯任意映射，因为在这种情况下不存在一个基础结构，所以概括只会引起讨厌的干扰。然而，即使对于这种任务，分布式表述也可以变得相当有效，并且在它们受到破坏时，会显示出某些心理学上的有趣影响。

在分布式表述能够得到有效使用之前，有几个困难问题

必须加以解决。一个难题是确定用来表述一个条目的活动模式。所选模式与其他现存模式之间的相似性,将决定出现的概括和干扰的种类。寻找好模式以供使用,等价于寻找这个领域的基本规律性。这一学习问题在第Ⅱ部分的章节中讨论。

另一个难题是,弄清分布式表述与人工智能中使用的种种技术,如图式或层级结构描述之间的关系。现存的人工智能程序,在快速找寻最适合于当前状况的图式方面,遇到很大困难。在迅速地把大量知识应用于这种最佳配合搜索方面,并行网络提供了潜力,但是只有当存在一个在并行网络中实现图式的好方法时,这种潜力才被获得。关于怎样可以做到这一点的讨论,见第 14 章。①

参考书目

Anderson, J. A. (1977). 'Neural Models with Cognitive Implications.' In D. LaBerge and S. J. Samuels (eds.), *Basic Processes in Reading Perception and Comprehension*, pp. 27–90. Hillsdale, NJ: Erlbaum.

Bartlett, F. C. (1932). *Remembering*. Cambridge: Cambridge University Press.

Collins, A. M., and Loftus, E. A. (1975). 'A Spreading-Activation Theory of Semantic Processing.' *Psychological Review* 82: 407–25.

Coltheart, M., Patterson, K., and Marshall, J. C. (1980). *Deep Dyslexia*. London: Routledge & Kegan.

Fahlman, S. E. (1979). *NETL: A System for Representing and Using Real-World*

---

① 本章以第一作者的一篇技术报告为基础,这一研究得到"系统发展研究基金"的资助。感谢 J·安德森,D·阿克利,D·巴拉德,F·克里克,S·法尔曼,J·费尔德曼,K·隆盖-希金斯,D·诺尔曼,T·塞诺斯基和 T·沙利斯与我们作了有益的讨论。

Knowledge. Cambridge, Mass.: MIT Press.
—— (1980). *The Hashnet Interconnection Scheme*. Tech. Rep. CMU-CS-80-125. Pittsburgh: Carnegie-Mellon University, Department of Computer Science.
Feldman, J. A. (1982). 'Dynamic Connections in Neural Networks.' *Biol. Cybernetics* 46: 27–39.
Hinton, G. E. (1981). 'Implementing Semantic Networks in Parallel Hardware.' In G. E. Hinton and J. A. Anderson (eds.), *Parallel Models of Associative Memory*, pp. 161–88. Hillsdale, NJ: Erlbaum.
—— and Anderson, J. A. (eds.) (1981). *Parallel Models of Associative Memory*. Hillsdale, NJ: Erlbaum.
Hopfield, J. J. (1984). 'Neurons with Graded Response Have Collective Computational Properties Like Those of Two-State Neurons.' *Proc. Nat. Acad. Sci.* (USA) 81: 3088–92.
Kawamoto, A. H., and Anderson J. A. (1984). 'Lexical Access Using a Neural Network.' *Proc. Sixth Annual Conference of the Cognitive Science Society*, 204–13.
Levin, J. A. (1976). *Proteus: An Activation Framework for Cognitive Process Models*. Tech. Rep. No. ISI/WP-2. Marina del Rey, Calif. University of Southern California, Information Sciences Institute.
Luria, A. R. (1973). *The Working Brain*. London: Penguin.
McClelland, J. L. (1981). 'Retrieving General and Specific Information from Stored Knowledge of Specifics.' *Proc. Third Annual Meeting of the Cognitive Science Society*, 170–2.
Minsky, M., and Papert, S. (1969). *Perceptrons*. Cambridge, Mass.: MIT Press.
Neisser, U. (1981). 'John Dean's Memory: A Case-Study.' *Cognition* 9: 1–22.
Newell, A. (1980). 'Physical Symbol Systems.' *Cognitive Science* 4: 135–83.
Norman, D. A., and Bobrow, D. G. (1979). 'Descriptions: An Intermediate Stage in Memory Retrieval.' *Cognitive Psychology* 11: 107–23.
Quillian, M. R. (1968). 'Semantic Memory.' In M. Minsky (ed.), *Semantic Information Processing*, pp. 227–70. Cambridge, Mass.: MIT Press.
Willshaw, D. J. (1981). 'Holography, Associative Memory, and Inductive Generalization.' In G. E. Hinton and J. A. Anderson (eds.), *Parallel Models of Associative Memory*, pp. 83–104. Hillsdale, NJ: Erlbaum.
Wood, C. C. (1978). 'Variations on a Theme by Lashley: Lesion Experiments on the Neural Model of Anderson, Silverstein, Ritz, & Jones.' *Psychological Review* 85: 582–91.

12 联结论、语言能力和解释方式

A·克拉克*

1. 解释层次和等价类概念

解释看来是一个多层次的事物。单个现象可以归入一套越来越一般的解释图式。解释往往有所得也有所失，我们用较低层次上详尽的描述/解释力度，来换取较高层次上对适用广度的满足。而在每一个这样的层次上，都存在着优点和缺点。某些解释可能只在某一层次上才有效，但是这样归类的个别情况也会发生变动，变动的方式只有沿着解释的一般性阶梯下降才能得到说明。

举例来说，达尔文或新达尔文的自然选择理论被置于很高的一般性层次。它描绘了一些非常一般的情况，在这些情况下，“盲目”选择能够产生出看起来是目的论的(或有目的的)进化演变。这一奇迹的出现，需要的是根据适应性而建立的特异性繁殖和某种向后代传递种性的机制。这是极具一般性和说服力的思想。因此这种最高层次解释的优点在于，它涵盖了一个由众多事例构成的无界集合，其中所

包含的实际**机制**(如传递机制)可能很不相同。这样,它就定义了一个由许多机制构成的**等价集合**,即这些机制可能在许多方面都全然不同,但因能够满足达尔文的要求而统一起来。

当然,优点往往伴随着缺点,达尔文的一般性解释的缺点也是显而易见的。在任何给定情况下,我们都尚未弄清达尔文的要求是**怎样**满足的。也就是说,对于任何给定情况下的遗传率和传递的实际机制,我们连最模糊的概念都没有。此外,显然还会有许多事实涉及某个特殊类别的情况(例如孟德尔豌豆中的隐性特征),这些事实是一般达尔文理论所未曾预见到的,因此我们有理由进一步寻找更加专门和详尽的解释。

孟德尔遗传学提供的正是这种解释。它从理论上假定了一组控制着每一特征的实体(现在称为基因),并描述了这些实体相结合的必然方式,以解释观察到的与相继各代豆株进化有关的各种事实。例如,这种说明中包含基因对(基因型)的思想,在基因对中有一个基因可能处于支配地位,因而解释了有关隐性特征的事实(有关进化论和孟德尔遗传学的通俗说明见 Ridley 1985)。

我们可以顺便指出,在任何两个层次(例如达尔文和孟德尔的遗传学)之间,几乎总是存在着理论上不容忽视的其他层次。所以孟德尔的遗传说事实上只是一个称做魏斯曼遗传说

---

* A·克拉克的"联结论、语言能力和解释方式"将登载在待出版的一期《英国科学哲学杂志》上。作者允许重印。

A·克拉克(Andy Clark),苏塞克斯大学认知与计算科学学院认知研究室哲学讲师。——译者

更一般机制的一个特例(见 Ridley 1985:23)。但是较之达尔文的遗传说,魏斯曼遗传说的一般性更低。魏斯曼学说从一般的达尔文事例中分割出一个理论上统一的子集合。而孟德尔学说又从魏斯曼学说中分割出一个理论上统一的子集合。在每一阶段上,等价类(equilavence class)都被策略地重新定义,从而将大量以前的成员排除在外。我们可以把这种情况想象为等价类的大小在逐渐收缩,尽管这不是严格如实的,因为每一新类别都可能有无穷多的成员,所以我猜想它们的大小是相同的!

孟德尔遗传学之所以提供了一个有意义的情况,还有一个原因。它最初被认为是对较低层次的以 DNA 为基础的遗传(即遗传机制的硬件实现方式)细节作出了简洁说明。正如丹尼特指出的,孟德尔基因被看作是对"以 DNA 块的方式直接实现的遗传语言"的说明(Dennett 1988b:385)。这相当于我们在谈到认知科学(语言能力理论)的某一抽象理论层次与实际加工策略之间关系的经典主义看法时将要用到的术语。但是事实上按照丹尼特的说法:

> 在"豆袋遗传学"的语言与分子细节及发展细节之间,有着理论上的重大悬殊,这些悬殊严重到足以表明,在考虑到所有事物的情况下,根本不出现(按照经典式理解的)基因(Dennett 1988b:385)。

这看起来像是(丹尼特也认为是)在模拟经典语言能力理论的结构体中命定的联结论观点。

即便当前的观点如此,也有可能在孟德尔遗传学层次之

下还存在着某个更深层的物理实现方式(上帝知道这两者之间有什么,如前所述),这样就完成了我们沿着解释的一般性阶梯的下行过程。我们从一般性的达尔文理论定义的一个由各种具体机制组成的大而多变的等价类的最高层次(层次-1)开始。我们下降到对一个由多机制子类的更详尽的说明(孟德尔理论),然后又以这种或那种方式到达 DNA 中那些机制的实施细节。其结果类似于一个三角形,它建立在地球动物遗传的真实细节之上,而这些遗传细节来自宽广得多的、支配着全部可能领域的集合的解释原理(见图 12-1)。

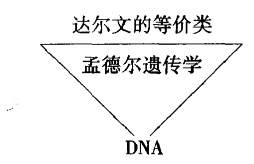


图 12-1 解释的一般性的得与失。以 DNA 为基础的描述在详尽力度方面所获得的,正是在跨界范围方面所损失的。

根据玛尔、乔姆斯基,以及纽厄尔和西蒙,还有其他人建立的概念,认知科学中的解释方式也有一个类似的多层结构。对已知类,或任务类(如视觉、语法分析等等),存在着一个以"计算的是什么和为什么计算的抽象概括"为内容的最高层次的描述,一个对执行计算的特殊算法的较低层次的说明,以及一个(更低层次的)关于算法怎样由物理硬件来实现的解释。玛尔以富里埃分析为例,对这一点作了具体说明。在最高层次上我们有着关于富里埃分析的一般概念,它可以通过几种不同的算法来实现。同时每一算法又可由许多不同种类的硬

件组织形式来完成(见 Marr 1977:129)。

在最高层次(玛尔称之为层次-1)的"正式"解释与给出"层次-1"理论的实际做法之间有着显著的差距。因为虽然正式路线是层次-1解释只对计算的**是什么**和**为什么**计算作出的说明,但是这一说明可以逐步改进,定义一个信息含量更高(即限制性更强)的等价类。层次-1理论的更精致的形式(它还达不到完整的算法解释程度),已由C·皮科克以"层次-1.5"为题作了有说服力的辩护(见 Peacoke 1986)。

皮科克强调的对照存在于两个等价类之间,一个是通过定义**外延式**函数(即由函数的结果,也就是玛尔所说的"**是什么**")生成的等价类,一个是限制性更强的(信息含量更高的)、通过说明用于一种算法的**信息体**而生成的等价类。因此,我们改写一下皮科克自己的一个例子,假定计算目标是,根据视网膜上的尺寸R,计算深度D和实际尺寸P。此外假定这一计算是在D、P和R值的一个限定范围中进行的。函数的外延式说明仅仅告诉我们,一旦给定某个D和P作为系统的输入,它就会产生某个确定的R作为输出。完成该任务的方法之一(我这里改写一下M·戴维斯所用的技巧(见 Davies 即将出版的著作,以及下文§2)),是对应于D、P值的每一组合,存储一组合法的R值——一个简单的查寻表。第二种方法是根据方程P=D×R来处理数据。我们说系统利用P=D×R这一信息时,正如皮科克所认为的,我们所做的要**多于**函数的外延式说明。因为查寻表**并不**利用这一信息,但依然属于由函数的外延式说明所生成的等价类。然而我们所做的又**少**于对一个特殊算法的说明,因为计算该方程的方法有许多种(例如采用不同算法做乘法)。

在本文以下部分中，当谈到**语言能力理论**时，我想到的就是这一点儿分析(即皮科克所说的层次-1.5)。这看来至少与乔姆斯基的做法是一致的，他造了“语言能力理论”这一术语，用来说明他本人对语言知识结构所作的有特色的研究的深度。它很可能与玛尔在“层次-1”的实际做法相一致，虽然与正式的规定不一致。

乔姆斯基的语言能力理论远远超出了对函数的外延式说明。它其实是在寻找对“什么构成了关于语言的知识”这一问题(在从大脑物理机制和从特定算法得出的抽象层次上)的回答。这种做法是在寻找一个“为可被掌握的人类语言所共有的原理和元素的框架”(Chomsky 1986:3)。同时(根据其最新的说法)把这一框架刻划成一个很独特的“与某些变动参数相联系的原理系统和一个带有它自己的若干分量的标记系统”(Chomsky 1986:221)。就本文的目的而言，重要的不是乔姆斯基实际提出了什么原理和参数。我们倒是只应该注意，如果一个语言能力理论与乔姆斯基模型一样地确定和有结构(毕竟他是这样说的!)，那么它就更像层次-1.5分析，而不是像简单的层次-1解释。因为它在某一抽象层次上描述了一种加工形式的结构(通过对加工所利用的信息作出说明)，所以它有助于“指导对机制的寻找”(Chomsky 1986:221)。简言之，这更像孟德尔的遗传学，而不是像一般的达尔文学说。它不仅仅是对一类结果的**描述**，而且还是对以我们作为其成员的一类机制的加工结构的**提示**。语言能力理论(至少如我们目前对它们所了解的)是否真正对人类加工的形式具有提示作用，正是本文的主题。

2. 经典的上下贯通形式

这样,语言能力理论就扮演着双重角色。它既说明了要计算的函数,同时又说明了某类算法所用的知识体或信息体。在经典认知科学中,这两种职能很容易同时得到履行。因为语言能力理论只是在符号数据结构上定义的一组联合起来的规则和原理。既然经典认知科学有赖于符号加工的构造体系,自然就要(在层次-2上)直接表述数据结构(如句子的结构描述),然后借助(在语言能力理论中)被定义规则和原理的显式的或默认的表述形式完成加工过程,以便在那些结构上运作。这样,由词干加词尾组成的变形动词的结构描述一经给出,经典主义者就可以进而定义层次-2的计算过程,即用词干加上-ed构成过去时(或任何别的时态)。于是经典主义者(借助于使用符号加工构造体系,实现层次-2算法)得到了唯一恰当的位置,使得在语言能力理论与它的层次-2实现过程之间保持一种十分密切的关系。的确,人们会产生这种看法:这种密切的关系似乎就是经典方法中的**构成性**。于是丹尼特就把经典主义者的梦想看成包含着"一个成功穿越玛尔三个层次的上下贯通形式"(Cascade)(Dennett 1987:227)。在我看来,经典主义者的基本看法的这种特征,正好符合福多尔和佩里舒最近就经典/联结论划分所作的说明。

福多尔和佩里舒认为,在建立认知模型的真正联结论方法与经典方法之间有两个基本区别。(这里"真正的联结论"把那些用单元和联结的子结构来完成经典理论的情况排除在外。)这些区别是:

1．“经典理论假定了——而联结论理论没有假定——一种‘思维语言’。”

这就是说，经典理论假定了具有一定形式的心理表述方式（数据结构）。这种表述方式具有**句法结构**，即它们是系统地构成的，方法是把原子成分结合成分子组合体，然后再（在复杂情况下）构成整个数据结构。简单说，它们假定了句法和语义相结合的**符号系统**。

2．“在经典模型中，心理状态转换所依据的原理，或一个输入在选择相应输出时所依据的原理，是根据心理表述方式的结构特性来定义的。因为经典的心理**表述方式**具有组合式结构，所以就可以参照它们的形式将经典心理**运作**施于它们。”（引自 Fodor and Pylyshyn 1988：12—13）

这就是说，如果**已知**你有某种（语言式的）有效的结构表述方式（如第 1 点所要求的），就可以在这些表述方式上定义计算运作，这样，这些运作所敏感的正是这一结构。如果这结构不存在（即如果没有符号表述），你就无法做到这一点！（虽然你有可能通过确定一个恰当的外延式函数使得**看起来**似乎做到了这一点。）

总之，经典系统是这样一个系统，它假定了具有句法结构的符号**表述方式**，同时它又利用这种表述方式的结构定义了适用于它们的计算**运作方式**。

在任何这类情况下，计算运作方式都可以通过在句法结构表述方式之上定义的转变或推导规则加以描述。例如：

如果（A 和 B）那么（A）

如果（A 和 B）那么（B）

如果（词干 + 词尾）那么（词干 + -ed）

括号中的条目是结构描述，它们将选出一些经典表述的无界类。“如果-那么”是对运作方式的说明。但应注意，在遇到行为术语时，经典主义者并不采用**显式**表述“如果-那么”条款的系统。需要外显的只是运作所依据的结构描述。因此，机器可以成为硬连线的，以便采用(A 和 B)形式的表达式，并把它们转换为表达式(A)和(B)。这样，推导规则就可能是隐含的，即**默认的**；但是数据结构必须是显式的。在这一问题上，福多尔和佩里舒正确地坚持道：

> 经典计算机相对于它们的程序而言可以是**规则隐含的**……在经典计算机中，**确实**需要外显的，不是它的程序，而是写在磁带上(或是存储在寄存器上)的符号。然而，它们所对应的不是机器的状态转变规则，而是它的数据结构(Fodor and Pylyshyn 1988:61)。

作为例子，他们指出，由语言学理论所假定的语法，在经典计算机中无需显式表述。但是语法据以定义(例如通过动词词干和子从句等)的**句子的结构描述**必须显式表述。因此从语言学的语言能力理论到层次-2 加工过程陈述，一个成功的“经典上下贯通形式”能够容许将语法规则装入计算机中。这样就表明了，企图仅仅用规则是否是显式的作为刻划经典主义/联结论差别的根据，是错误的。

然而，目前存在的危险是，我们忽略了(经典主义者)怎样把语言能力理论(或推导规则和数据结构的集合)说成是与层次-2 实现过程有着密切关系的。因为我们说过(见上 § 1)，如果已知比如说像“P = D × R”这样的简单语言能力理论，就

不会再去建立一个只存储着某个有限论域的所有 P、D 和 R 的合法值的系统。然而这样的系统肯定有 P、D 和 R 的显式表述方式。所以如果没有，必然是由于它连推导规则"P = D × R"的默认知识也没有。于是问题就成为，我们怎样找到产生这一差异的原因？建立在像规则"P = D × R"那样的默认知识属性上的约束情况，是不需要显式表述的，但是又是哪些约束排除了作为那个规则的默认知识例子的查寻表了呢？当我们准备问（§3 和 §4）联结论系统是否具有关于经典规则的默认知识时，这个回答将是十分重要的。

M·戴维斯（根据 E·埃文斯提供的信息）提出了如下意见："要使一个说话者具有关于特定联合式理论的默认知识，说话者内部必须有一个反映这一理论中的推导结构的因果解释结构"（Davies，待出版：4）。按戴维斯的意思，"理论中的推导结构"是指转变规则（如 P = D × R）。那么，体现出一个"反映"这种推导结构的"因果解释结构"是怎么回事呢？根据戴维斯的看法，只是在更高层次上在每一个被认为是包含推导规则的例子所做的加工过程陈述中具有一个**因果公因子**（戴维斯的用语）。这样，对于查寻表的情形来说，在加工各种 P、D 和 R 值的**所有例子**中，不必存在任何因果公因子。反之，如果整个加工路线都经历的因果公因子是**存在**的，那么（在一些无关紧要的附带条件下，见 Davies 待出版，和 Davies 1987）完全可以说系统具有该规则的默认知识。如戴维斯所指出的，这一结果正好与我们的认知神经心理学的直觉相符合。因为系统若满足如此解释的默认知识的约束，自然有可能造成这样一种破坏：因果公因子的损害将使得解决整个一类问题（例如说明 P、D 和 R 的问题）的能力完全丧失。另一方面，那

些不能满足这种约束的系统，自然可能造成较少的系统损失（例如，查寻系统可能丧失关于 P、D 和 R 的某些合法组合的知识，但却保留了关于其他组合的知识）。类似看法也适用于过去时生成的情形。具有关于“取出词干，加上-ed”这一规则的默认知识的系统可能丧失所有形成规则过去时的能力。通过查寻完成这一任务的系统，则不会出现这种情况。

戴维斯的解释（将遁词虚拟“因果”公因子标准化，见下文§3）看来是令人信服的。将其加以归纳，我们就最后得出任何真正的经典认知模型都具有的下列特征：

（根据对语言能力理论的态度定义的经典主义。）

一个认知模型，如果它的加工层次描述与一个标准语言能力理论的结构有着某种相当密切的关系，那么它就是经典的。一个标准的语言能力理论假定了一组推导规则或原理，经定义，可按照它们的形式将它们应用于一类结构式的符号表述方式。根据要求，密切关系包括（1）在加工层次描述中，对定义规则所依据的结构表述方式的显式表述，以及（2）对那些规则和原理本身的显式或默认的表述。一个规则或原理只有在这种条件下才被判定是以默认形式表述的：在加工层次的描述中存在着一个因果公因子，无论该规则或原理在语言能力层次的转变说明中是否被援引，该因子都是在起作用。

这就是“经典上下贯通形式”的内容通过玛尔解释层次时所呈现的复杂曲折的细节（我对此表示抱歉）。联结论像水坝一样阻拦了这个“瀑布”[1]。它是怎样做到这一点的，又会产生什么样的渠道，这是我们在下文中要讨论的内容。

---

[1] “上下贯通形式”和“瀑布”同为 cascade，意为瀑布状物。——译者

3. 牛顿式语言能力

对于结构式语言能力理论与层次-2加工过程陈述之间的关系，联结论的看法与§2中设想的那种简洁的“上下贯通形式”是根本不同的。层次-2陈述不是被看作为对语言能力理论的推导形式的反映，而是被看作与它的一种关系，有点像牛顿力学与量子物理学之间的那种关系。物理宇宙实际上并不是牛顿的，但是在某些可说明的条件下，它的表现非常像是牛顿的。这样，在一定的事例范围内，牛顿原理就可以描述和预见物理系统的行为。但是在某种直觉的、然而有点难以捉摸的意义上，那些原理并没有描述决定着物理行为的实际作用力。鲁梅哈特和麦克莱兰十分欣赏这个类比，他们写道：

> 或许可以认为，常规的符号加工模型是宏观的说明，类似于牛顿力学，而我们的模型提供了更为微观的说明，类似于量子理论。由于对牛顿力学与量子论间关系的透彻理解，我们就能理解，宏观的描述层次可能**只是**对更加微观的理论的**近似**(Rumelhart and McClelland 1986a: i. 125)。

为了说明这一点，我们来看一看P·斯莫伦斯基提供的一个简单的例子。设想要建立模型的认知任务包含定性地回答与特定电路的行为有关的问题。(对单个电路作出限制，可能会把经典主义者吓一跳，尽管斯莫伦斯基为它辩护，根据是，

少量的这种表述的作用可能像多功能专门知识中使用的"组块"一样,见 Smolensky 1986: ii.241。)给出对电路的描述,专家就能回答像"如果我们提高 A 点的电阻,对电压会产生什么影响?"这样的问题(也就是电压升高,降低,或是保持不变)。

通常的方法是,我们假定:为这一任务制定算法所依据的信息是由高层次语言能力做出理论说明的,在这一说明的推导过程中引用了各种电路学定律(也就是斯莫伦斯基所谓的电路学"硬定律",欧姆定律和克希霍夫定律)。例如,与欧姆定律有关的推导要引用这一方程:

电压(V) = 电流(C) × 电阻(R)

在上文 § 2 中,我们只确认了层次-2 加工陈述可能与这种语言能力理论具有相当密切的关系的两种方式。在最简单的情况下,加工过程可能包含着用符号表述的欧姆定律,系统将它读出,并遵循它。在较为复杂的情况下,它可能包含有关欧姆定律的默认知识,在一组状态转变中通过一个因果公因子展现出来。(顺及:斯莫伦斯基本人在这里的做法,看来是不恰当地强调了简单选择方式,见 Fodor and Pylyshyn 1988; Pinker and Prince 1988; Davies 待出版;Clark 1989。)

在斯莫伦斯基提出的简单电路问题求解的联结论模型的情形中,没有一种上下贯通形式是起作用的。为了弄清其原因,我们需要看一看这一模型的形式。这个模型对电路状态的表述,是通过一组特性单元上的活动模式来完成的。它们对电路变元中出现的定性变化进行编码,即在训练实例中,它们对电阻 R1 增大时总电压的升降等情况进行编码。这些特性单元与一组被斯莫伦斯基称为"知识原子"的东西相联结,知识原子表述了分布在特性单元子组上的活动模式。事实上

它们是对电路学的实际定律所允许的特性单元状态的合法组合进行编码。例如"系统对欧姆定律的知识……分布在众多的知识原子上，这些知识原子的子模式对电流、电压和电阻的合法特性组合进行编码"(Smolensky 1988:19)。简言之，对于定性变化的每一合法组合，都存在着一个子模式(GS 子模式，或所讨论的电路的"知识原子")。

初看起来，这个系统似乎只是查寻表的单元和联结的一个实现过程。然而情况并非如此。事实上，联结论网络只有在配备了过量的隐蔽单元，从而能够对输入输出配对作简单记忆时，才具有查寻表的功能。相比之下，被这系统编码的是斯莫伦斯基称为"软约束"的东西，亦即通常存在于各种不同的特性单元(微特性)之间的关系模式。这样，它就具备了电路微特性之间定性关系的"一般知识"。但是它不具备压缩在像欧姆定律那样的硬约束中的一般知识。软约束是特性单元与知识原子之间的双向联结，这种联结以这种或那种方式对网络施加影响，但并不强迫它；也就是说，这种联结能被其他单元的活动压倒，这也是它之所以"软"的原因。正如在所有联结论网络中那样，系统所作的计算是由同时尝试满足尽可能多的这种软约束来完成的。为了弄清系统并不仅仅是合法组合的查寻树，我们只需注意到，对于那些在合法组合的简单查寻表中找不到答案的(有矛盾的或不完全的)问题，它有能力作出合理的回答。

软约束经数字编码，成为单元之间的加权联结强度。这样，问题求解就由"一系列多节点(即单元)最新校正数据"来完成，"每一最新校正数据都是在形式的数字规则和数字计算的基础上作出的微观决策"(Smolensky 1986:ii.246)。

网络的两个性质使我们特别感兴趣。第一,不难看出,**如果**给出一个提得恰当的问题,**并且**有无限的加工时间,那么它**总是**给出正确答案,和电路学硬定律所预言的一样。但是正如已经说过的那样,它并不局限于这些定律。给它一个提得不当或有矛盾的问题,它会满足尽可能多的软约束(它真正知道的就是这些)。这样,"在提得恰当的问题和无限加工时间的理想化区域之外,该系统就会具备合理的性能"(Smolensky 1988:20)。这样,就可以把(欧姆定律等)硬规则看作是理论家对系统实际性能的理想化子集所做的外部特征刻划(如果这使我们想起丹尼特关于"意向性态度"发表的意见,决不是偶然的,见 Dennett 1987)。

第二,网络在反复尝试满足所有的软约束时,表现出有趣的**系列**行为。斯莫伦斯基把这种系列行为说成是一组**宏观决策**,其中每一决策都相当于"部分网络承担部分解答任务"。斯莫伦斯基指出,这些宏观决策"类似于产生规则的启动过程。事实上,这些'产生'的'启动'顺序与符号正向链接推理系统中的顺序基本上是相同的"(Smolensky 1988:19)。这样看来,网络敏感的似乎是处在细粒度描述中的硬的符号规则。事情不会**简单**到它好像知道硬规则似地解决"外延式"问题。甚至问题求解的**各个阶段**也可能看起来像是因系统对语言能力理论获得的符号推导步骤进行加工模拟而造成的。

但是,根据前面 §2 中的说法,这种情况只是一种幻觉。该系统既没有关于硬规则的显式知识,也没有关于它的默认知识。其原因不难理解。很显然,它并没有以显式方式向自己表述欧姆定律。例如,并不存在简洁的单元子模式,可以被认为代表着出现在欧姆定律中的一般的理想电阻。相反,一

些单元组代表电阻 R1,另一些单元组代表电阻 R2。在比较复杂的网络中,那些在活动时代表像电阻这样的最高层次(或概念层次)概念的单元联合体,表现出高度的**语境敏感**。就是说,它们是随着事件的语境而变动的。因此,采用斯莫伦斯基本人的例子,**咖啡**在这种网络中的表述就不只是一个单一的、重复出现的句法条目,而是一些随语境而切换的较小条目(微特征)的联合体。在杯子的语境中,咖啡可由包含"液体""接触陶瓷"的联合体来表述。在罐子的语境中,咖啡可能包含"颗粒""接触玻璃"。这样看来,这里只有一个"贯穿诸语境的'咖啡向量'的近似等价物",它不像"符号加工系统中贯穿不同语境的咖啡标志的精确等价物"(Smolensky 1988:16)。用这样的方法将概念层次的符号"咖啡"代之以微特征的切换联合体,即所谓"维度切换",这样一些系统本身就丧失了既在经典**语言能力**理论中使用,**也在**经典符号加工(**层次**-2)解释中使用的结构式心理表述方式。同样,在被描述的简单网络中,也不存在代表电阻的稳定的表述实体(就像在名声不佳的过去时网络中不存在代表"动词词干"的稳定的、重复出现的实体一样,见 Rumelhart and McClelland 1986b; Pinker and Prince 1988; Clark 1989)。其直接后果是,对于可参照概念层次结构体的规则来说,并不存在显式表述。缺少**默认**表述几乎是其直接后果,因为对当时不存在的结构,加工过程的敏感性就无从谈起。

若以我们惯用的方式来看待这一问题,就不能说该系统**默认**地表述了那些规则,因为在它的问题求解中不存在因果公因子,使得例如在语言能力理论中,无论何时引用欧姆定律,那个单一的因子都在产生实际结果的加工过程中起着关

键的作用。为了弄清这一点，我们只须回想一下，在与 R1 的命运有关的问题的求解中，以及与 R2 的命运有关的问题的求解中，不同的特性单元和知识原子都将起关键作用。在这个(限定的)意义上，它的确与查寻树有某种共同之处。因为网络之所以不能体现关于规则的严格的默认知识，是由于它无法使整个实际加工过程按照一定路线通过一个与以反复引用欧姆定律为标记的推导瓶颈相对应的因果瓶颈。由于在语言能力理论具有单一推导方程的地方，网络具有多重因果路线，所以网络就无权获得关于规则的严格的默认知识。在这一点上，它不能体现关于规则的默认知识的原因是和查寻树一样的。

现在来讨论前面提到的通词。在我就我的理解采纳戴维斯对默认知识的特征刻划时，我为使用“因果公因子”这个短语感到不安。它的优点是使神经心理学的内在联系看起来非常直观，但是它也可能掩盖了层层叠加的虚拟机的某些复杂性。因为据我的猜测，要实现的经典上下贯通形式所需的共同的东西，与其说是一个简单的物理状态，不如说是一个**加工陈述定义于其上的虚拟机**的状态。说到底，即使是经典系统，继承了运作系统的各种长处，也不可能在它每次经历一个由欧姆定律(在语言能力理论中)标出的加工过程转变时都采用同样的**物理**状态。然而层次-2 加工描述不需要(也不应该)标示出这个区别，因为就实际算法而言，它没有任何内在联系，仅仅是一个实现过程细节而已。相反，在联结论陈述中可能对应于单一符号转变的各种不同的状态，**必然**在加工/算法描述中被标示出来。系统的真正知识终究还是这样编码的知识——这一事实是大大夸张了的联结论加工的流动性和语境

敏感性的直接原因。我不能肯定这一点造成的差别有多大，因为虚拟机同真实机器一样，可以表现出不同的分类模式，因而与认知神经心理学有密切关系。

把遁词放在一边，现在我们能够总结建立语言能力理论的牛顿式态度了。牛顿式联结论者会把语言能力理论看做对加工过程的**描述**(可能是很细的粒度——回想一下关于“宏观决策”的讨论)。但是他不会把该理论看做对有关的实际加工的**提示**。该理论之所以不是提示性的，是因为其行为并不依赖于系统所具有的关于符号推导规则的显式的或默认的知识；这一事实在提得恰当的问题和无限的加工时间构成的理想化的、“牛顿式”的领域之外的系统行为中得到证实。这一行为表明，“该理论确实始终是一个‘量子’系统”(Smolensky 1988:20)。

这一观点是在一个有启发性的脚注中(Smolensky 1986: ii.246)提出的，其提法十分适合于我们的讨论。斯莫伦斯基指出，把语言能力的特性刻划成一组应用于符号系统的推导规则，这可以看作是为生成系统的高度协调(= 软约束最大限度地被满足)的状态提供了一种**语法**。这样，语言能力理论就作为一套定律而出现，这些定律的作用是挑选出系统在某些理想条件下所处的状态。所以这就是对待语言能力理论的完全的牛顿式态度：一个语言能力理论是一种确定出系统某些稳定状态的语法。因此在主要的事例范围内它具有描述上的恰当性。但是它并没有揭示出斯莫伦斯基所说的系统的**动力学**，或是系统的实际加工策略。它并不具有导向层次-2加工陈述的恰当的提示作用。因此对持牛顿立场的人来说，建立语言能力理论已失去它原来的意义了。

4. 次等语言能力

因此，根据牛顿式联结论模型，在有点理想化的事例范围内，语言能力理论就具有描述上恰当的导向输出的功能。然而，对联结论者来说这并不是对语言能力理论的唯一理解。的确，这种理解不同于隐含在对高层次问题求解的另一些联结论处理方法中的那种理解。本节中，我打算讨论另一类不同的处理方法，我把它们称做语言能力的**次等**(rogue)模型。

简单说，牛顿模型与次等模型之间的基本差别是这样的：在牛顿模型中，联结论网络**本身**能在理想化条件下，以语言能力理论规定的所有方式起作用。与之呈对照的是，在次等模型中，基本的联结论网络**本身**(即使加工时间和提得恰当的问题都处于理想状态时)并没有能力产生出语言能力理论所要求的(即可从中推导出的)全部结果。其实，应该看到，就人类实际上表现出全面的经典语言能力而论，他们能这样做只是因为利用了**其他资源**(例如，用来操纵符号的连接式符号加工或现实世界结构——像笔和纸)。这样，由次等方法产生的对语言能力模型的看法就是：在这一领域中的快速日常问题求解中它们有暂时使用并非在线的附加资源的情况。

在鲁梅哈特、斯莫伦斯基、麦克莱兰和欣顿(1986)的著作中有一个次级模型的例子。该例子同我们做乘法运算的能力有关。我们可以设想这里有一个用到算术定律的符号语言能力理论。但是基本的联结论模型不同于这样一种符号存储

器。更确切地说,它是一个训练有素的模式匹配器,它能立刻“看出”某个乘法运算的结果。例如,我们大都能“看出”7×7的答案,但“看”不出 7984×5431 的答案。那么我们是怎样解决后一类问题的呢?

推测是这样的:“答案来自我们创造人工制品的能力,即我们有这样一种能力,能创造出一些有形的表述方式,使我们能以简单方式进行操作,以获得对十分困难和抽象的问题的解答”(Rumelhart, Smolensky, McClelland, and Hinton 1986: ii. 44)。于是,为了求解 7984×5431,我们可以把这个问题写出来,**然后**小心地运用一系列我们擅长的简单模式匹配步骤来解决它,例如从 4 与 1 相乘开始,继续做下去。

7984

5431

——

······4

他们接着说,我们甚至可以学会**在我们的头脑中**完成这一运算,方法是把这种外部符号换成我们内部的某种表述方式。但是在本质上它仍然是一个由我们操作的“外部”符号工具,同时它仍然构成一种资源,这种资源是建立在我们利用的基本联结论模式匹配能力的**最高层次之上**的。(最近 D·丹尼特也就语句似乎遍及我们头脑的一些事例发表了非常类似的看法。在这些事例中,我们的确是在做经典符号加工。但是这种加工也可能构成附加资源,这是不包含在我们所有的日常、非语言推理过程中的,见 Dennett 1987:233,114 - 15;也见 Clark 1988。)

对多位数乘法所作的解释,当然还存在许多疑问,因为整

个过程似乎包含着对于支配连续运用模式匹配能力的符号规则的了解！但是我们已经看到，很多表面上依赖于符号的行为，有可能在符号以下的层次上产生。（详细的探讨请看 Clark 1989。）无论如何，我只是用这例子就这种会构成次等模型的解释表明了一种态度。

作为最后一个例子（这应归功于 M·戴维斯），我们来考察我们分析歧义句子的能力，如"The horse raced past the barn fell."[①]。作语法分析的次等模型可能会遇上这样的事情。我们有一个在线的迅速而粗略的联结论网络，它能对我们在日常谈话中遇到的大多数句子作语法分析。但是它没有分析歧义句子的能力（甚至原则上受制于理想方式）。然而我们还有一个（不是在线的，而是在背景中的）经典符号分析器（是不是有点像一个 ATN?），它能够分析这种句子。当迅速而粗略的网络失败时，这个备用者就出来挽回局面。这符合现象学，初看起来好像无意义的句子，后来变得清楚了。在这情形中，经典语言能力理论正确地描述了备用系统的结构。然而它没有描述在线的网络。此外，如果我们设想，经典的备用系统在训练网络的过程中是活动的，那么在一系列简单的实例上，这两个系统达到局部汇合，就不会令人惊奇了。

次等方法所具有的明显而有关联的优点，涉及到所谓监督学习算法在心理学上的可取性。有一些步骤用来训练这样一些联结论网络，它们依赖于错误信息的反向传播，因而依赖于一台教导机（通常是常规计算机），它注视着系统的输出，并

---

① 这句可解释为"经过谷仓的赛马倒下了"，或"马儿疾驰过有谷仓的山岗"。——译者

告诉系统这输出**应该**是怎样的。(稍详细一点的内容见§5中对NET**讲话**的讨论。)这种安排常常在心理上显得很不合实际。例如,在学习语言时,我们只能根据给出的**正面**例子来学习(像乔姆斯基论者热衷于指出的那样)。那么,教导机和错误信息又是从何而来的呢?

次等模型开辟了一种可能性,这就是由一个单独的系统存储一组输入输出配对(例如一组观察到的印刷符-音位配对),并用它们来训练联结论网络。因此负面的例子是由大脑本身,而不是其他执行者生成和确定的。T·塞诺斯基最近也表示支持这种说法,并引用白顶雀的例子加以说明。白顶雀听到父鸟唱歌以后,要等到明年才会唱。他的假设是,这只鸟以某种方式把歌**存储**起来,但是必须训练一个使它再现的网络——这一过程解释了呈现与再现之间的一大段空白。回到我们的主题,次等方法显然为联结论学习的反向传播法在心理学上受到尊重提供了最大的希望。

在最极端的情况下,次等模型有可能使人类的在线加工与严格的语言能力模型相分离,而使经典语言能力**恢复**为完整、恰当的备用系统描述。要注意,经典语言能力理论在次等模型上所处的地位,与它在牛顿式模型上的地位显然不同。对建立次等模型的人来说,经典语言能力理论恰当地描述了一个虽然不是固定地在线的、但却是重要的加工系统类别。事实上我觉得,这些经典资源的重要性,即使是在那些口头上承认了它们存在的地方,也还没有被完全领会。所以斯莫伦斯基(Smolensky 1988)提出一个思想,把语言看作一种知识传递的特殊手段,它含有一个加工过程,是由一个被称为意识规则解释程序的经典虚拟机完成的。但是语言指令

的作用仍然以稍居第二等的性质存在着。有了语言，我们才能建立规则，例如在训练的初始阶段，这些规则对新手会有帮助（也见 Smolensky 1986：ii.251－2，文中把基本上相同的描绘用于前面讨论过的电路问题求解的例子）。然而，**专家**被描绘成使用了一个强有力的联结论网络，似乎只要有语言，就可以把他的见解的封装元件传递给他人。这可能大大低估了符号加工所作的贡献。这种加工也可以提供一种根据专家自己的在线推理而形成的元反射，以帮助专家理解和扩展他自己的技能。（一些有关假设见 Karmiloff-Smith 1987 和 Dennett 1988a。）

采用次等方法造成的最有影响的后果，是使当前流行的关于心灵的"正确"认知结构的辩论大大地复杂化了（见 Fodor and Pylyshyn 1988）。因为如果采用一个次等模型，这些问题就没有唯一的答案。任何有关人类认知技能的圆满解释都需要用到**两种**模型，而经典形式不只是一种方便的近似。似乎物理世界原来在一些领域中是牛顿的，在另一些领域是量子的，而不是一致地可用量子来描述，仅仅在某些情况下**看来像**牛顿的。

总括起来说，次等模型甚至否认了经典语言能力模型对于在线加工的**描述上的**恰当性。但是它们承认经典理论对于**附加资源系统**的加工来说，既是描述性的，**也**是提示性的。这一附加资源系统保证了在给定领域中可称为人类典范推理能力的东西。在次等场合，语言能力模型是原来的模型（对某个加工策略的准确描述），但是它不在原来的**地方**，因为它不描述日常在线加工的计算形式。

5. 联结论解释的方法论

看来,联结论解释策略**不能**适合于纽厄尔和西蒙提出的模子。联结论者不能以纽厄尔和西蒙式的语言能力理论作为开始,然后简单地在层次-2算法模型中实现它。在我们看来,其原因是简单明了的。这种语言能力理论是由一组转变规则组成的,这些规则被限定用于数据结构的标准符号表述方式。在经典模型中,这些数据结构在机器中以显式方式表述(经典的函数构造**恰恰**就是使之为可能的那种构造)。于是机器就依照这些规则来操作它们(规则**本身**在任何这种数据结构中都不必以显式标示)。形成对照的是,在特殊的联结论模型中,并没有正好与经典符号数据结构相对应的东西。相反地,语境敏感的、切换的单元联合体对应的是单个的经典表述方式。这就是前面提到的维度切换。既然不存在对经典符号结构简洁的模拟,该系统就**不能**(即使是默认地)体现出**正是**在**这些结构**之上定义的关于转变规则的知识。所以经典语言能力理论不能充分地提示联结论层次-2的加工过程陈述。假如它具有提示作用,"联结论"系统就只不过相当于快速而果断地实现**经典**认知模型(见 Fodor and Pylyshyn 1988)。

从以上内容中我们看到了,在有关经典语言能力模型的两种立场中,虔诚的联结论者只能接受其中之一。这两种立场就是前面讨论过的牛顿式立场和次等立场。但是较深入的、基本的问题仍未解决。因为,在否认任何最高层次经典语言能力理论能够充分提示层次-2加工策略这一点上,牛顿式立场和次等立场是一致的,这里的层次-2加工策略是关于完

成给定认知任务的中央在线联结论网络的。但是现在看来，这一点(见前文§1)是一个令人加倍难堪的损失，因为经典语言能力理论执行着双重任务。首先，它描绘出认知科学研究所取的恰当形式(即勾勒出在形成语言能力理论层次上的任务的轮廓，然后编写算法来完成它)。第二，它描绘出认知科学中的**解释**包含些什么内容。拥有一个工作程序，仅这件事本身不能看作是对我们怎样完成一个给定的认知任务作出了解释。相反，我们需要的是关于程序遇到什么约束以及为什么必然会遇到这些约束的某种最高层次的理解——这一理解是通过给出最高层次语言能力理论而自然提供的，该理论可以看作是由一个给定的程序**类**来实现的。这样，经典语言能力理论失效时，就预示着这种危险：使联结论模型在相当深层的意义上成为**非解释性**的。同时也给联接论研究的实际**方法论**留下了疑难。

作为对该问题简单的说明，让我们来看一个例子——认知科学中有效的老式解释(GOFEICS，引自J·豪格兰)。以朴素物理学为例。众所周知，朴素物理学试图找到一种知识，使活动的、具体存在的事物能设法找到绕过复杂的物理世界的道路。这个一般性方案的一个著名的例子就是海斯在液体的朴素物理学方面所作的研究(Hayes 1985b)。该工作包括尝试编一个"液体所能具有的各种可能状态的分类"，以及制定一组与运动、变化和液体几何形状有关的规则。最后的理论包括对液体15种状态的说明，以及在谓词演算中开列的74条编号的规则或公理。这相当于一个详尽的语言能力说明，最终可能得出它的完整的层次-2算法形式。的确，海斯十分明确地主张对这一方案要作高层次的研究，他坚持认为过早地

寻求工作程序是错误的(见 Hayes 1985a:3)。这样,朴素物理学的解释策略就是纽厄尔和西蒙所推荐的正规的经典方法论的范例。首先,寻找一个包含符号表述和一组状态转变规则的高层次语言能力理论;然后,编写出实现该语言能力理论的层次-2算法;我们对于算法所满足的必要条件具有精确的、较高层次的理解,因而真正地领会了算法为什么能够完成当前任务,由于这样一些知识,这种算法是可靠的。这种可靠性正是联结论所缺少的,因为它的运行方式不是(**不能够**)通过系统阐述一个详细的经典语言能力理论,然后再根据经典的符号加工构造简洁地实现这一理论。

于是就面临这样的问题:联结论者**应当**怎样行事,什么构成了对于加工过程的较高层次的理解,这些就是我们为了断言真正**解释**了一个任务是如何完成的所需要了解的东西。看来,我们需要的是对经典语言能力理论解释层次的某种联结论模拟。

我相信这种模拟是存在的。但是只有我们在认知科学中对于我们关于解释的描述发动一场哥白尼式的革命,它才会显露其面目。因为联结论者成功地翻转了这一惯用的时间和方法论的解释顺序,这与哥白尼用地球围绕太阳转动来代替另一种转动方式,从而翻转了那个时代惯用的天文学模型是非常相像的。同样,在联结论理论的建立中,是使高层次理解围绕着一个已经学会怎样对付某个认知领地的工作程序在转动。这样就翻转了正规的玛尔式顺序,根据玛尔式顺序,高层次理解(即语言能力理论)是首要的,并严密地指导着算法的求得。为了弄清这一点,同时也为了知道联结论者的高层次理论是怎样背离经典语言能力理论形式的,我打算考察一下

塞诺斯基的 NET 讲话方案。

NET 讲话是一个大的分布式联结论模型，其目的是研究把书写输入（即单词）转换为音位输出（即声音或言语）的部分过程。网络构造包含一组每次由文本中七个字母激励的输入单元、一组隐蔽单元和一组为音位编码的输出单元。输出被送入口声合成器后，产生出实际的讲话声音。

网络从隐蔽单元（在选定参数内）的加权值和联结的随机分布开始，就是说它对于从文本到音位的转换规则没有任何“想法”。它的任务是通过重复呈现训练示例来学习对付这一特别棘手的认知领域的方法（这领域之所以棘手，是因为从文本到音位的转换是无规则的、次规则的和语境敏感的）。学习按标准方式进行，即按照反向传播的学习规则进行。其工作方式是，给系统一个输入，检查其输出（由一个计算机的“监督器”自动完成），并告诉它**应当**产生什么输出（即什么音位编码）。然后，学习规则使系统对隐蔽单元上的加权值作出微小的调整，使输出趋向正确。这一过程要重复数千次。奇妙的情况出现了，这个系统以听得见的方式慢慢地学会英语文本的发音，从咿哑学语，到半能辨认单词，再到相当可靠的最终演示。更充分的说明见罗森堡和塞诺斯基（Rosenberg and Sejnowski 1987），以及塞诺斯基和罗森堡（Sejnowski and Rosenberg 1986）的文章。

现在来看 NET 讲话方案的方法论。第一步无疑是引用这一领域的相当丰富的某种先验分析的成果。这一点表现在作者对输入表述方式的选择（例如，选择七个字母作为窗口，为字母和标点符号选择某种编码），对输出表述方式的选择（音位编码），以及对隐蔽单元构造（如隐蔽单元的个数）和学习规则的选择上。这些选择强调了联结论模型建立中的先验任务分析

在某种程度上继续发挥着重要作用。但是它们还远不是任何一种充分清晰表达的、从文本到音位转换的语言能力理论。因为显而易见,这里缺少任何一组定义在输入和输出表述上的专用的状态转变规则。于是,系统就面临着学习任务,要学习了解一组在其隐蔽单元上的加权值,使该加权值完成向期望状态转变的中介任务。由于这一原因,我认为联结论者的特征是从层次-0.5的"任务分析"开始其研究,而不是层次-1(或1.5)的语言能力理论。然而值得一提的是,层次-0.5的说明虽然不及充分发展的符号语言能力理论,仍然可以体现先验信息在心理学上的非现实部分。因为当一个人学会完成一个任务时,他事先并不知道要调用多少隐蔽单元(太多的话,就会形成一个不提供信息的"查寻树",太少的话,就无法处理数据),也不知道表述该解答的最好方式是什么。在这意义上,层次-0.5说明在问题求解方面所做的工作,比起某些联结论者愿意承认的,可能更多一些。然而对当前的目的来说,主要问题仅仅是,层次-0.5模型构成一个基础,在这基础上,由于有力的联结论学习规则,系统(经过大量训练之后)变得能够对付目标中的认知领地。从这一点来看,联结论者已经掌握了一个工作系统——一个全面的层次-3实现方式。

假定就此停步。我们就在增进对本文-音位转换现象的理解的方式方面得到了一个有用的玩具,但是是一个很小的玩具。当然,联结论者并不就此停步,他们现在必须做的工作是,从大有发展前途的层次-3的实现方式回到对这一任务的较高层次的理解上。这就是玛尔透镜。这一较高层次的理解怎样获得呢?

可使用的策略是各种各样的,将来还会发现更多。这里

我只说三种。第一，简明的、非微观层次的**观察**。给出一个特定的输入，联结论者就可以看到(隐蔽单元中)所产生的单元活动模式。(网络在一台能记录这种活动的常规计算机上被模拟时，情况总是这样的。)正如塞诺斯基所指出的，这样就提供了神经科学家不得不费力去搜集的那种数据。因为在记录单一细胞的活动方面，神经科学具有精良的技术。但是在记录分布于大量细胞上的同步活动的模式方面，它并不处于有利的地位。(也见待出版的 Churchland 1989。)第二，**网络病理学**。故意损坏人的大脑，以帮助我们弄清细胞的次级组合体在各种不同任务中所起的作用，显然是违反道德标准的，然而损坏人工神经网络，看起来要容易接受得多。最后，也许是最重要的，联结论者能够向我们描绘系统是以何种方式学会对试图要对付的认知空间进行划分的。这种描绘是由所谓"分层簇分析"完成的，在我看来，正是这一描绘为高层次的、语言能力理论的理解提供了最接近于联结论的模拟。

"何种表述方式已被编码在网络的隐蔽单元之中?"这就是簇分析要回答的问题。这是一个难题，因为如前所述，表述方式一般具有较为复杂的、不明显的、维度切换的特征。为了弄清簇分析是如何工作的，我们把网络的任务看做是配置隐蔽单元的加权值，其方式是使它能完成一种集合的分割。其目标是使隐蔽单元，当且仅当该输入正好应该得到这一独特的输出时，以独特的方式作出响应。这样，在从文本到音位的转换中，我们希望隐蔽单元在以"the"作为输入时的表现与以"said"作为输入时会有很大不同。但是我们希望它们在以"sail"和"sale"作为输入时的表现是等同的。所以隐蔽单元的任务就是以与当前工作相适应的方式分割一个空间(由这些单元的数目和它们

可能的激活水平所确定)。一个非常简单的系统,如丘奇兰提出的岩石/矿藏网络(Churchland 1989, 待出版),可能只需把由隐蔽单元所确定的空间分割为两个主要子空间——一个独特的模式用于标明矿藏的输入,一个用于标明岩石的输入。为使文本-音位转换的复杂性保持原样,NET讲话必须把它的隐蔽单元空间划分得更细小(事实上分割得使79种可能的从字母到音位的配对中的每一种有一个独特的模式)。正如罗森堡和塞诺斯基(Rosenberg and Sejnowski 1987)所做的那样,簇分析实际上在隐蔽单元激活的79种独特的稳定模式的这一基本层次之上构造了一个层级分割。把79种模式一一取出,并把它和最接近的相邻者即与它共同之处最多者配对,就构成这一层级体系。这些配对用作下一步分析的建造模块,在下一步中,(原有那对的成员之间的)一个平均激活轮廓图就被计算出来,并和它的最近的相邻者配对,这个相邻者是从通过对原有每对取平均值而生成的二次图形库中取得的。反复进行这一过程,直到生成最后的一对。这就是网络学到的最粗略的隐蔽单元空间划分——在NET讲话的情形中,这一划分其实是对应于元音和辅音之间的划分的。

这样,簇分析就为可能的隐蔽单元激活的空间形状提供了一种描绘,而隐蔽单元的激活加强了网络的性能。通过对这个空间各个方面(即各种成簇现象)的考虑,理论家可望得到对该系统当前行为的某种深入认识。例如,可能发现这系统原来是对某个一直未被注意到或是被当做不重要的次级规律性高度敏感的。这就如同在我们试图理解认知空间时,为我们提供了对该空间的形状的勾画。这种勾画必须加以解释,而这是一个实际的、有时是困难的任务。但这并不是暗中

摸索,因为我们能弄清什么输入同那种构形有联系(即使这是一个由簇分析揭示出的较高层次的构形)。

这样我们就得到了当前每一类别的成员——于是该任务就是寻找用来描述各类别成员资格条件的概念层次上的明白易懂的术语。

我想指出,一个得到充分解释的簇分析构成了最接近于经典语言能力理论的联结论模拟。像语言能力理论一样,它提供了一个比算法层次更高(即更一般)的理解层次。因为在联结论者看来,"算法"说明的内容必须包括(a)网络构形,(b)单元规则和联结强度。但是在这种算法说明和特定的簇分析之间存在着多对一的映射,例如,一个在开始时带有一组不同的随机加权值的网络,经过训练,就会表现出**相同**的分割轮廓图(因而具有等同的簇分析),但这样做用的是一组差别很大的个体加权值。然而,簇分析与经典语言能力理论不同,通常它看来**不**像一组定义在概念层次实体上的状态转变规则。相反,它更像一块认知领地的形状的某种几何图像。那些认为高层次解释必须像一组命题和规则的理论家,可能会感到难以适应这一点。

另一方面,一些激进的反命题的理论家(如 Churchland 1989, 待出版)或许会认为,解释后的簇分析给予日常陈述话语的东西**太多**了。丘奇兰的看法是,正确的理解层次位于联结的加权值层次。他坚持认为,原因在于这些就是系统"真正"知道的一切。系统并没有表述它自己的诸分割。此外,在给定新输入时,两个系统的学习方式可以不同,即使它们在时刻 t_1 的簇分析是等同的。因为联结的加权值(我们已了解到它与簇分析的关系是多对一的)是认知进化中的关键性部分。

然而，这看上去就像常见的高层次解释那种有所得也有所失的情况。有时我们要选择一个分析层次，使之对特定联结的加权值的说明进行分组，而成为由公共簇分析支配的等价类，这时我们自然地以特殊性换取了一般性。正像纯粹的达尔文主义未对隐性特征加以解释，而是强调了存在于一类进化机制中的一般原则一样，簇分析也未对认知发展的某些细节作出解释，而是强调了能使一类网络设法成功地对付给定认知领地的总的感受性。如果联结论准备对认知性能做深入的解释，那么某个像这样的高层次理解看来就是非常重要的。仅仅只对一组联结的加权值作出说明，当然不是一种**解释方式**，甚至连反命题论者也是这样看。

然而，我想强调的主要论点与对簇分析优缺点的任何看法都不相干。它关系到传统认知科学的方法论转换。不管用什么手段，联结论者是通过对一个**学会**完成当前任务的网络的反思和修修补补，来获得对一个认知任务的高层次理解的。与受玛尔鼓舞的经典理论家不同，联结论者不是先制定好（句子、符号的）语言能力理论的骨架，然后再赋予它算法的血肉。相反地，它从层次-0.5开始，训练一个网络，然后再尝试掌握已由网络体现出的高层次规则。可以说这给认知科学带来了不可思议的好处，因为该学科一直被**特别规定**和命题主义的（有关的）弊病困扰着。由于不得不将语言能力理论归结为一组组定义在经典的符号数据结构上的规则，理论家们已经凭空抽取出一些原理，以便使他们的工作条理化。与之不同的是，联结论方法论允许该任务提出自行勾画轮廓的要求，从而，以不受标准符号公式化要求左右的方式提出该空间的形状。于是我们就避免了把我们有意识的命题思想的形式强加

于我们无意识加工的模型之上——这种强加方式一般地说既在实践上是不成功的，也在演化上显得异乎寻常。

总之，联结论者不得不在没有经典语言能力理论援助的情况下设法做事情，但是它既没有失去高层次解释力量，也没有失去方法论的合理性。相反地，联结论解释的方法论完全做到了避免专门的组织原则，及避免命题、语言的偏颇。还剩下几个重要的、尚未解决的问题，它们与联结论可能提供的提取和表达这种高层次解释的最好方法有关。但是像簇分析、网络症状研究和激活记录这样一些技术业已形成，毫无疑问将会得到充分理解。一旦如此，认知解释中的哥白尼革命就走上坦途了。

6. 结论：瀑布、堤坝和分流

实际的加工(层次-2)模型与传统语言能力理论应有什么样的关系，对这一问题的看法，经典主义者和联结论者看来必然存在着根本分歧。本文在开始时列举了经典主义者对这一关系的看法，以及两种联结论的替代方案。可以方便地把它们描绘如下。

关系一：瀑　　布

丹尼特把经典主义者的观点说成是一个"成功穿越玛尔三个层次的瀑布"的观点(Dennett 1987:227)。一旦经典符号加工构造体系存在，瀑布就会顺畅地流动。语言能力理论

的公理或规则是用语言表达的公式,可用来从一个符号推导出另一个符号。这样,各种(层次-2)算法就可以实现这一推导结构。它们可以以显式方式(通过标示出规则)或默认方式(依照规则通过加工显式符号串)来完成。根据这一经典观点,层次-2是对层次-1的简洁的回应。

关系二:堤　　坝

牛顿式联结论拦截住经典的瀑布,其方法是引入层次-1推导规则运作于其上的条目(符号串)与联结论网络"运作于其上的"条目(亚符号)之间的维度切换。层次-1理论可对网络行为的某些(理想化的)方面加以描述。但是网络既不体现推导规则的显式知识或默认知识,也不体现这些规则得以定义的概念层次的结构。

关系三:分　　流

次等模型表述了一种实际性能取决于两个系统的较复杂的事态。一个是日常在线系统,适用于牛顿式联结论者所描述的那种语言能力理论,即它同某些内蕴行为相匹配,但未体现出经典知识。另一个是附加资源,可利用外部符号来建立,它是对经典机器的模拟。它本身能够体现出语言能力理论中所说明的推导规则和概念层次结构。次等模型提出许多相互作用的(虚拟)构造,使得关于"正确的"认知构造的辩论更加复杂了。

因此联结论者必须以某种方式使自己与经典语言能力模型的细节保持距离。对于在线的联结论加工形式,这种模型不具有恰当的提示作用,虽然它们可能对这种加工的结果(的

一个子集合)或另外某个认知资源具有描述作用。但是这种联结论和语言能力理论的错位,引起了一个严重的问题。因为经典主义方法论保证了对于已建立模型的认知现象作出有用而准确的较高层次的理解。联结论者却不同,可能看起来有一些工作系统,但是对这些系统并没有较高层次的理解,所以在一定意义上没有对认知现象作出解释。

一旦我们打算对认知科学中我们关于解释的思考进行一场哥白尼式革命的时候,上述忧虑就会减轻。在玛尔的影响下,认知科学家们倾向于期望对先于算法编写和贯穿于算法编写的任务作出某种高层次的理解。这正是经典语言能力理论说明致力于完成的工作。然而联结论者成功地反转了这一策略。他们从对任务的最低限度的理解开始,训练网络去完成这个任务,然后用各种原理方法寻求获得有关网络正在做什么和为什么这样做的较高层次的理解。这可能包括仔细记录网络的活动状况,检验网络在受到各种不同形式损害后的行为,和查明网络的隐蔽单元对它们所对付的空间进行划分的方式。这最后一步(文中已提到的簇分析)显然提供了一种较高层次的理解,因为在已知的簇分析与那组能够实现它的联结加权值之间,存在着多对一的关系。联结论者从层次-0.5 模型开始,迅速地发展到层次-3 实现方式,然后必然返回到详细的较高层次理解。

我想指出的是,这一解释方式的转换,实际上构成了联结论方法超过传统认知科学的一个主要优点。它之所以是优点,是因为它提供了一种方法,可避免以专门方式生成公理和原理。在使某个认知任务条理化时(使我们想起朴素物理学),联结论者不必根据一组相当任意的符号式的、以语言为

基础的公理作出决定，而可以让任务自行组织网络，然后才尝试概括出对它的活动的各种较高层次的描绘。此外，这些描绘可能背离（其方式还需要我们去充分想象）把理论作为一组命题的传统描绘方式。相反，它们可能具有更多的几何或图形的性质，或是以意想不到的、看起来很笨拙的方式使用语言（Churchland 1989，待出版）。

最后，我们可以推测：在对待语言能力的态度上，联结论者和经典主义者之所以分化，是有其相当深刻的原因的。这原因就是语言能力模型是一个用命题或逻辑形式表示的传统理论。经典主义者认为思维只是对具有命题或逻辑形式的条目的操作；而联结论者则坚持认为，这只不过是一种表面现象，思维（"深层"思维，而不仅仅是列举命题）取决于对多种十分不同的结构的操作。结果是，经典主义者试图给出一个层次-2加工模型，这模型正是定义在与它的层次-1理论描绘完全相同的各种结构之上。而联结论者则坚持要分解这一结构，并用某种全然不同的东西来替代它。

这里出现了一个奇怪的、具有讽刺意味的情况。在人工智能的早期，战斗口号是"计算机不是单调地处理数字，它们是操作符号！"其用意是以表明计算与思维多么相像来鼓舞持怀疑态度的公众。现在轮子转了整整一圈。联结论系统的优点看来就在于"它们不是操作符号，它们是单调地处理数字"。而今天，我们都知道（难道不是吗？）思维不仅仅是符号操作！轮子仍在转动着[①]。

---

① 本文的工作是法英哲学和认知科学合作研究项目的一部分。我特别感谢M·戴维斯，他就§4中语言能力次等模型的思想同我进行了交谈，并提出了建议，我也特别感谢M·博登和T·塞诺斯基，他们对§5中的内容进行了讨论。

参考书目

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origin and Use.* New York: Praeger Publishers.

Churchland, P. (forthcoming: 1989). 'On the Nature of Theories: A Neurocomputational Perspective.' In P. M. Churchland (ed.), *The Neurocomputational Perspective.* Cambridge, Mass.: MIT Press.

Clark, A. (1988). 'Thoughts, Sentences and Cognitive Science.' *Philosophical Psychology* 1: 263–78.

—— (1989). *Microcognition: Philosophy, Cognitive Science and Parallel Distributed Processing.* Cambridge, Mass.: MIT/Bradford Books.

Davies, M. (1987). 'Tacit Knowledge and Semantic Theory: Can a Five Per Cent Difference Matter?' *Mind* 96: 441–62.

—— (forthcoming). 'Connectionism, Modularity and Tacit Knowledge.' In *British Journal for the Philosophy of Science.*

Dennett, D. (1987). *The Intentional Stance.* Cambridge, Mass.: MIT/Bradford Books.

—— (1988*a*). 'The Evolution of Consciousness.' Jacobsen Lecture, University of London, May 1988. *Tufts University Current Circulating Manuscript* CCM-88-1.

—— (1988*b*). 'Review of Psychosemantics.' *Journal of Philosophy* 85: 384–9.

Fodor, J., and Pylyshyn, Z. (1988). 'Connectionism and Cognitive Architecture.' *Cognition* 28: 3–71.

Hayes, P. J. (1985*a*). 'The Second Naïve Physics Manifesto.' In J. R. Hobbs and R. C. Moore (ed.), *Formal Theories of the Commonsense World*, pp. 1–36. Norwood, NJ: Ablex.

—— (1985*b*). 'The Ontology of Liquids.' In J. R. Hobbs and R. C. Moore (eds.), *Formal Theories of the Commonsense World*, pp. 71–108. Norwood, NJ: Ablex.

Karmiloff-Smith, A. (1987). 'Beyond Modularity: A Developmental Perspective on Human Consciousness.' Transcript of talk given to the Annual Meeting of the British Psychological Society, Sussex, April 1987.

Marr, D. (1977). 'Artificial Intelligence—A Personal View.' Repr. in J. Haugeland (ed.), *Mind Design*, pp. 37–47. Cambridge, Mass.: MIT/Bradford Books.

Peacocke, C. (1986). 'Explanation in Computational Psychology: Language, Perception and Level 1.5.' *Mind and Language*, 1 (2): 101–23.

Pinker, A., and Prince, S. (1988). 'On Language and Connectionism.' *Cognition* 28: 73–193.

Ridley, M. (1985). *The Problems of Evolution.* Oxford: Oxford University Press.

Rosenberg, C., and Sejnowski, T. (1987). 'Parallel Networks That Learn to Pronounce English Text.' *Complex Systems* 1: 145–68.

Rumelhart, D., and McClelland, J. (1986*a*), 'PDP Models and General Issues in Cognitive Science.' In McClelland, Rumelhart, and the PDP Research Group (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1, pp. 110–46. Cambridge, Mass.: MIT/Bradford Books.

—— (1986*b*), 'On Learning the Past Tenses of English Verbs.' In McClelland, Rumelhart, and the PDP Research Groups (eds.), *Parallel Distributed Proces-*

sing: Explorations in the Microstructure of Cognition, Vol. 2, pp. 216–27. Cambridge, Mass.: MIT/Bradford Books.

Rumelhart, D., Smolensky, P., McClelland, J., and Hinton, G. (1986). 'Schemata and Sequential Thought Processes in PDP Models'. In McClelland, Rumelhart, and the PDP Research Group (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 2, pp. 7–57. Cambridge, Mass.: MIT/Bradford Books.

Sejnowski, T., and Rosenberg, C. (1986). 'NETtalk: A Parallel Network That Learns to Read Aloud.' John Hopkins University Electrical Engineering and Computer Science Technical Report, JHU/EEC-86/01.

Smolensky, P. (1986). 'Information Processing in Dymanical Systems: Foundations of Harmony Theory.' In McClelland, Rumelhart, and the PDP Research Group (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 2, pp. 194–281. Cambridge, Mass.: MIT/Bradford Books.

—— (1988). 'On the Proper Treatment of Connectionism.' *Behavioral and Brain Sciences* 11: 1–74.

13 造就心灵还是建立大脑模型：人工智能的分歧点

H·L·德雷福斯和 S·E·德雷福斯*

在我看来，最大的可能就是有一天人们会达成这种明确见解：在……神经系统中并不存在与**特定**思想，或是**特定**观念，或是记忆相对应的摹本。

L·维特根斯坦(Wittgenstein 1948: i. 504(66e))

信息并不是存储在什么特定的地方，而是存储在各个地方。与其说信息是"发现的"不如说是"召来的"。

D·鲁梅哈特和 D·诺尔曼

(Rumelhart and D. Norman 1981: 3)

50年代初期，当计算机器正在盛行起来的时候，一些前驱思想家们就开始意识到：数字计算机所能做的不仅仅是嘎吱嘎吱地咬嚼数字。那时在可以把计算机看作什么的问题上，出现了两种对立的观点，每一观点都有与之相联系的研究纲领，它们使出浑身解数，以期得到承认。一派把计算机看作操作思想符号的系统；另一派则把计算机看作建立大脑模型的手段。一派试图用计算机来例示对世界的形式表述；另一派则试图用计算机模拟神经元的相互作用。一派把问题求解作为智能的

范式,另一派则把学习作为智能的范式。一派利用逻辑学,另一派则利用统计学。在学派上,一个是哲学中理性主义、还原论传统的继承者;另一个则把自己看作理想化的、整体论的神经科学。

前一派的战斗口号是:心灵和数字计算机都是物理符号系统。1955 年,A·纽厄尔和 H·西蒙在兰德公司工作时,已经得出这样的结论:由数字计算机操作的二进制数字串能代表任何东西——数字当然没有问题,此外还有现实世界中的许多特征。而且可以用程序作规则来表述这些符号之间的关系,使系统能进一步推断出有关所表述物体和它们的关系的事实。正如纽厄尔最近论述有关 AI 争论历史时所指出的:

> 数字计算机的工作范围决定了计算机是操作数字的机器。拥护此观点的人认为,重要的事情是每一样东西都可以经编码成为数字,指令也不例外。反之,AI 科学家们把计算机看作操作符号的机器。他们认为,重要的是每一样东西都可以经编码成为符号,数字也不例外(Newell 1983:196)。

这种看待计算机的方式成为看待心灵的方式的基础。纽厄尔和西蒙假设道:人类大脑和数字计算机尽管在结构和机制上

---

* H·L·德雷福斯和 S·E·德雷福斯所著"造就心灵还是建立大脑模型:人工智能的分歧点"一文引自《人工智能》117,no.1(1988 年冬季号),马萨诸塞州剑桥美国艺术和科学研究院杂志《代达罗斯》允许重印。

S·德雷福斯(Stuart Dreyfus)加利福尼亚大学(柏克莱)工业工程与运筹学系教授。

H·德雷福斯(Hubert Dreyfus)加利福尼亚大学(柏克莱)哲学教授。

全然不同，但是在某一抽象层次上具有共同的功能描述。在这一层次上，人类大脑和恰当编程的数字计算机可被看作同一类装置的两个不同的特例，这一装置通过用形式规则操作符号来生成智能行为。纽厄尔和西蒙把他们的观点陈述为一个假设：

> **物理符号系统假设**。对于一般智能行为来说，物理符号系统具有的手段既是必要的，也是充分的。
>
> 所谓“必要”的意思是：任何表现出一般智能的系统，经过分析，都可以证明是一个物理符号系统。所谓“充分”的意思是：任何足够大的物理符号系统，都可以通过进一步地组织，而表现出一般智能（Newell and Simon 1981:41）。

纽厄尔和西蒙将其假设的来源追溯到G·弗雷格、B·罗素和A·N·怀特海（Newell and Simon 1981:42），然而弗雷格及其追随者们自身无疑也继承了一个悠久的、原子论的理性主义传统。笛卡尔已作过这样的假定：所有的理解都是由形成和操作恰当的表述方式组成的，这些表述方式可以经分析成为本原元素（**原创性单质**），一切现象都可以理解为这些简单元素的复杂结合形式。另外，与此同时，霍布斯也隐含地假定，这些元素是由纯句法运作来陈述的形式分量，所以推理可以还原为计算。霍布斯写道：“一个人在**推理**时，他所做的只不过是将许多小部分相加而构造出一个整量，因为**推理**……不是别的，就是计算……”（Hobbes 1958:45）。最后，莱布尼兹在致力于建立经典的数学观念——每一事物的形式化——的研究时，曾寻求支撑物来发展一个普适的符号系统，使“我们能

为每一客体指定一个确定的特征数"(Leibniz 1951:18)。根据莱布尼兹的看法,在理解过程中,我们把概念分析成更简单的元素。为了避免向越来越简单的元素回归,必须有终极的单质,据此可以对所有复杂概念作出理解。此外,如果要将概念应用于现实,必须要有这些元素所代表的简单特征。莱布尼兹设想出"一种人类思想的字母表"(1951:20),它们的"字符必须在被用于证明时显示出客体中也发现了的某种联结、分组和排序"(1951:10)。

在弗雷格和罗素的基础上,L·维特根斯坦在他的《逻辑哲学论》中阐述了心灵与实体关系的这种句法的、表述的观点的纯理论形式。他把世界定义为逻辑上独立的原子事实的总和:

> 1.1　世界是事实的总和,而不是事物的总和。

他认为,事实转而也可以彻底地分析为本原客体。

> 2.01　原子事实是一些客体的结合……
>
> 2.0124　如果所有客体都为已知,则所有原子事实亦因而为已知。

维特根斯坦主张,这些事实,以及它们的组分和它们的逻辑关系都在心灵中得到表述。

> 2.1　我们为自己建造了事实的形象。
>
> 2.15　形象的元素以确定的方式相互结合,从而表述了事物也是这样相互结合的(Wittgenstein 1960)。

我们可以这样认为:AI 就是试图找到主体(人或计算机)中的这些本原元素和逻辑关系,该主体映射出构成世界的本原客体和它们之间的关系。纽厄尔和西蒙的物理符号系统假设事实上是把维特根斯坦的看法(它本身就是经典的理性主义哲学传统的顶点)转换成一个经验主义的主张,并以此为基础形成研究纲领。

来自反面的直觉认识是:我们应当以建立大脑模型,而不是建立心灵关于世界的符号表述的模型的方式,来创造人工智能,这种认识不是受到哲学的启发,而是受到不久后被称为神经科学的思想的启发。这一启发直接来自 D·O·赫布的工作,他在 1949 年提出,如果当神经元 A 和神经元 B 同时被刺激时,该刺激使它们之间的联结强度增高,那么一团神经元就能够进行学习。

F·罗森布拉特继承了这一先见,他推理道,因为智能行为是以我们对世界的表述为基础的,所以它很可能是难于形式化的。AI 应当换一种方式,尝试自动完成这样的过程,即神经元网络用以学会恰当地辨别模式和作出响应的过程。正如罗森布拉特所指出的:

> [关于操作符号研究纲领的]隐含假定表明,对我们希望系统完成的行为作出详细说明,是相对容易的,而难点是设计出一个有效完成这一行为的装置或机制。……使**物理系统**公理化,然后用分析的方式研究该系统,以确定出它的行为,比起使**行为**公理化,然后用逻辑综合技术来设计物理系统,既是比较容易的,也是更为有益的

(Rosenblatt 1962b:386)。

指出这两个研究纲领的不同之处的另一方法是:寻求符号表述方式的人们是想找出一个形式结构,以使计算机具备解决某一类问题或是区分某些类型的模式的能力;而另一方面,罗森布拉特想要建造一个物理装置,或是在数字计算机上模拟这一装置,然后由该装置生成自己的能力:

> 我们已听说的许多讨论过的模型都同这一问题有关:如果要使一个系统表现出某个特性 X,该系统必须具备何种逻辑结构。这在本质上是一个有关静态系统的问题……
>
> 这个问题也可以换一种方法来看:何种系统能使特性 X **逐步形成**? 我想,我们可以通过若干有趣的实例来说明,即使第一个问题没有答案,第二个问题也可以得到解决(Rosenblatt 1962b:387)。

这两种方法都很快获得惊人的成功。1956 年纽厄尔和西蒙已经成功地用符号表述方式来解决简单的智力测验问题和证明命题演算中的定理,为计算机编程。有这些早期令人瞩目的结果作为基础,物理符号系统假设看来似乎就要被证实了,纽厄尔和西蒙也处于可以理解的志得意满之中。西蒙声称:

> 我并不是有意让你感到惊讶或震惊……然而我可以用最简单的方式作出总结:目前世界上存在着一些会思考、会学习、会创造的机器。并且它们做这些事情的能力正在迅速提高,在不远的将来,它们处理问题的范围,在

时空上将达到人类心灵已被应用到的范围(Newell and Simon 1958:6)。

他和纽厄尔解释道:

> 我们现在掌握了启发式的(不同于算法式的)问题求解理论的要素,而且我们使用这一理论,既可以理解人类的启发式过程,也可以用数字计算机模拟这些过程。直觉、顿悟、学习不再为人类所专有,任何大型高速计算机都可以通过编程表现出这些能力(Newell and Simon 1958:6)。[①]

罗森布拉特在一种他称为感知机的装置上将他的想法付诸实现。[②] 1956年时,罗森布拉特就能够训练一台感知机将

① 启发式规则是那些在人类使用时被看作为基于经验或判断的规则。这些规则常常引出一些问题的可取的解答,或是提高问题求解过程的效率。而算法保证了在有限时间内得到一个正确的解(如果解存在的话),启发法仅仅增加找出可取的解答的可能性。

② 鲁梅哈特和麦克莱兰(Rumelhart and McClelland 1986)对感知机有如下描述:"构成这样一些机器的是通常称为**视网膜**的东西,即一排有时被看作排列在二维空间布局上的二元输入;一组**谓词**,即一组二元阈值单元,它们与视网膜中的单元子集合具有固定联结,使每一谓词都计算出与它相联结的单元子集合上的某个局部功能;以及一个或多个决策单元,它们与谓词的联结是可修正的"(i.111)。他们对比了像感知机那样的并行分布式处理(PDP)模型存储信息的方法和符号表述存储信息的方法的不同:"在大多数模型中,知识作为一个模式的静态摹本被存储着。检索相当于找出长时记忆中的模式,并把它复制到一个缓冲存储器或工作记忆中。长时记忆中的存储表述方式与工作记忆中的活动表述方式并无真实区别。然而在PDP模型中并非如此。在这些模型中,被存储的不是模式本身,而是使这些模式得以被再创造的单元之间的**联结强度**(i.31)……关于任何个别模式的知识,不是存储在为这一模式而保留的专门单元的联结中,而是分布在大量加工单元之间的联结之上"(i.33)。这一关于表述方式的新观念,直接导致了罗森布拉特的这个思想:"这样的机器应该能够通过学习,而不是通过按特征和规则进行编程来获取它们的能力,因为如果知识存在于联结强度之中,那么学习要做的事必然就是找到正确的联结强度,这样就可以在正确环境中产生出正确的激活模式。对这一类模型来说,这是极为重要的特性,因为它提供了信息加工机制能够通过调节其联结方式而学会获得它在加工过程中所处的激活状态之间的"(i.32)相互依存性的可能性。

某些类型的模式分类为相似的，并把它们与另一些不相似的模式区分开来。1959 年时，他再次感到欢欣鼓舞，认为他的方法已得到证明：

> 看来很清楚的是……感知机引入了一种新的信息加工自动装置：我们第一次有了一台能够具备原创思想的机器。作为对生物大脑的模拟，感知机，更确切地说是统计可分离性理论，比起以前提出的任何系统，似乎更接近于满足对神经系统功能解释的要求……在概念上看来，感知机无疑已经建立起有可能体现人类认知功能的非人类系统的可行性和原理……根据统计学原理，而不是根据逻辑原理运作的信息加工装置的前景仿佛清晰地展现出来了（Rosenblatt 1958：i.449）。

60 年代早期，这两种方法看起来同样大有前途，同时也因言过其实的论断而同样使自己易受攻击。然而两个研究纲领之间这场内战的结局却出乎意料地发生了倾斜。到 1970 年，以感知机为范式的大脑模拟研究渐受冷落，资金不足，而主张用数字计算机进行符号操作的一方，却无可争辩地控制了资金来源、学位授予、杂志和专题讨论会，使他们的研究纲领呈现出生气勃勃的景象。

由于每个正在进行的研究纲领都制造出命定论的神话，重现这一变化是怎样发生的，是件颇为复杂的事情。在胜利者看来，似乎是符号信息加工获胜了，因为它选择了一条正确途径；而神经网络或联结论方法失败了，因为它一点效果也没有。但是关于该领域历史的这一看法，是一种缺乏发展眼光

的错误认识。两种研究纲领都有值得探究的思想，两者也都有深层的、尚未被认识的问题。

每一种立场都有其诋毁者，诋毁者们说的基本上是同样的话：每一种方法都表明它能够解决某些简易的问题，但是没有理由认为任一派别可以将它的方法推广到现实世界的复杂状况中去。的确，有证据表明，随着问题变得更复杂，两种方法所需要的计算是以指数方式增长的，因而很快就变得难以驾驭。1969 年，M·明斯基和 S·帕佩尔特对罗森布拉特的感知机发表看法如下：

> 罗森布拉特的方案很快生根，不久就有几乎多达上百个大大小小的小组用该模型做实验……这数百个项目和实验得出的结果，一般来说是令人失望的，其解释不能令人信服。这些机器在非常简单的问题上通常工作得不错，但是当分配给它们的任务变难时，情况就迅速恶化。(Minsky and Papert 1969)。

三年后，詹姆士·莱特希尔爵士在回顾了使用如西蒙和明斯基程序那样的启发式程序所做的工作之后，也得出十分类似的否定性结论：

> 大多数从事 AI 及其相关领域研究的人都承认，过去的 25 年中所取得的成就显然是令人失望的。研究者们在 1950 年前后甚至在 1960 年前后进入这一领域时，是抱着很大希望的，这些希望与 1972 年时的实际状况相去甚远。该领域迄今在任何一方面作出的发现都没有产生出当初指望的那种重大影响……

造成这些业已经受的失望的一个颇为一般性的原因是：没有认识到"组合激增"的牵连关系。对于构造一个以大量知识为基础的……系统来说，这是一个一般性障碍，它源于随着该基础规模的增长，任何组合表达式都具有的激烈增长，它们代表着众多的各种可能的根据特定规则对该知识基础的元素进行分组的方式。

正如D·鲁梅哈特和D·齐普泽对此所做的简明总结："组合激增或迟或早总会让你碰上，尽管在并行和串行中的方式有时会有所不同"(Rumelhart and McClelland 1986: i.158)。正如J·福多尔曾指出的那样，双方已走进一局三维象棋游戏，认为它是一场"连城"[①]游戏。既然如此，为什么在游戏的如此早的阶段上，在知道得如此之少，要学的如此之多的情况下，一组研究者全胜了另一组研究者呢？为什么在这一关键的分歧点上，符号表述方案成了"城中"仅有的棋局呢？

每一位了解该领域历史的人，都能八九不离十地指出其原因。1965年前后，明斯基和帕佩尔特正领导着麻省理工学院的一个专门用于符号操作方法研究的实验室，因而为赢得支持而与感知机方案展开着竞争，这时他们开始散发一本攻击关于感知机思想的书稿。在书中，他们阐明了自己的科学立场：

感知机作为"模式识别"或"学习"机器，受到了公众的广泛注意，并在大量的书籍、杂志文章和卷帙浩繁的"报告"中被讨论过。这些著作中的大部分……是没有科

① 一种类似于中国五子棋的游戏。——译者

学价值的(Minsky and Papert 1969:4)。

然而他们的抨击也是一场哲学运动。他们正确地看到,传统的依赖于向逻辑原素还原的方式正在受到新的整体论的挑战:

> 当前的作者们(起先是分别地,而后是共同地)都变得多少带有一点治疗强迫性:驱除掉我们害怕会成为早先的"整体论"或"格式塔"错误观念阴影的东西,因为它们很可能造成对工程和人工智能领域的纠缠,就像它们先前纠缠生物学和心理学一样(1969:19)。

他们是相当正确的。人工神经网络可以但并非必须允许根据人类能够识别和用来解决问题的特征对它的隐蔽节点[①]作出解释。虽然神经网络建模本身不对任何观点有所承诺,但是可以证明:联想并不要求隐蔽节点必须是可以解释的。整体论者像罗森布拉特乐意于假定单个节点或节点模式不是在挑选领域的固定特征。

明斯基和帕佩尔特是如此坚决地排除所有竞争,又如此笃信从笛卡尔直到早期维特根斯坦所保持的原子论传统,以致他们著作中的提法远远超出了实际的证明。他们试图分析单层感知机的能力,[②] 然而他们书中的数学部分完全忽视了罗森布拉特关于多层机的章节,以及他对以误差的反向

① 隐蔽节点是指这样的节点:它们既不能直接检测网络的输入,也不能构成网络的输出。但是通过具有可调整强度的联结,它们直接或间接地与检测输入和构成输出的节点连接起来。

② 单层网络没有隐蔽节点,而多层网络肯定含有隐蔽节点。

传播[1] 为基础的概率学习算法的收敛性的证明(Rosenblatt 1962a:292)。[2]鲁梅哈特和麦克莱兰指出:

> 明斯基和帕佩尔特试图说明何种函数是(单层)机器能够计算的,何种是不能计算的。他们特意证明了,如果不运用数量大到不可思议的谓词,这种感知机就不能计算像奇偶性那样的数学函数(无论视网膜上存在着奇数个点还是偶数个点)或是连通性的拓扑函数(无论其上所有的点是否直接地或通过也在其上的别的点与其上另外所有的点相联结)。这个分析是极其漂亮的,同时也证明了数学方法对于分析计算系统的重要性(Rumelhart and McClelland 1986: i.111)。

但是这一分析的意义有很大的局限性。鲁梅哈特和麦克莱兰继续写道:

> 基本的情况是,……虽然明斯基和帕佩尔特对**单层感知机**所作的分析是完全正确的,但是这些定理不能应用于甚至稍微复杂一点的系统。特别是,这种分析不能用于多层系统或是允许有反馈回路的系统(1986: i.112)。

然而在对**感知机**所作的结论中,当明斯基和帕佩尔特向

---

① 误差反向传播要求从输出节点开始,按递归方式计算出改变联结强度对期望输出与由输入产生的输出之间的差异的影响。然后在学习过程中对加权值进行调整,以减小这个差异。

② 参阅:"加上第四层信号传递单元,或使三层感知机的 A 单元交叉耦合,有可能在任意变换群之上得出概括问题的解答。……在反向耦合感知机中,可能出现对一个复杂区域内的熟悉对象的选择性注意。这样的感知机也可能有选择地注意那些相对于它们的背景作不同运动的对象。"(Rosenblatt 1962a:576)

自己提出了“你考虑过多层感知机吗?”这个问题时,他们在口头上承认该问题尚未解决,而给人的印象却是已经解决了:

是的,我们已经考虑了 Gamba 机,它可以被看作“双层感知机”。我们尚未发现(通过思考或是通过研究文献)任何其他一类真正使人感兴趣的多层机,至少那种具有一些看起来与感知机的原理有重要关系的……原理的机器从未发现过。我们认为,对于我们有关增加层数不起作用的直觉判断作出阐释(或是加以驳斥),将成为重要的研究课题(Minsky and Papert 1969: 231 - 2)。

他们对 AI 中格式塔思维所作的攻击,其成功的程度已超出他们的梦想。只有引不起注意的少数几个人还在探究这一“重要的研究课题”,其中有 S·格罗斯贝格、J·A·安德森和 T·科霍宁。的确,差不多所有从事 AI 的人都认为神经网络已经永远被搁置起来了。鲁梅哈特和麦克莱兰指出:

明斯基和帕佩尔特关于单层感知机局限性的分析,加上符号加工方法在人工智能中的某些早期成就,足以向这一领域中众多的研究者表明,对人工智能和认知心理学来说,感知机式的计算装置是没有前途的(Rumelhart and McClelland 1986:i.112)。

但是为什么“足以表明”呢? 两种方法都产生了某种有希望的成果,也都作出了某些缺乏根据的许诺。[1] 中止对任一

[1] 关于对直到 1978 年时符号表述方法的实际成绩作出的评价,参阅德雷福斯(Dreyfus 1979)。

种方法的考虑，都为时过早。然而明斯基和帕佩尔特著作中的某些东西撞中了一根敏感的弦。AI 研究者们似乎都对激起这一攻击的整体论抱有类似宗教的哲学偏见。人们可以在例如纽厄尔和西蒙关于物理符号系统的文章中看到这一传统的力量。他们的文章从心灵和计算机因操作离散符号而成为智能的这一科学假设作为开始，但是却以这样一个揭示作为结束："有关逻辑和计算机的研究已经向我们揭示，智能存在于物理符号系统之中。"(Newell and Simon 1981:64)

整体论无法与这样一些强烈的哲学信念相匹敌。罗森布拉特以及数百个受到他的工作的鼓励而从事网络研究的责任较小的小组都名声扫地了。他的研究经费枯竭了，他的著作的出版也遇到了麻烦。到了 1970 年，对 AI 来说，神经网络已经销声匿迹。纽厄尔说，在他的 AI 的历史上，符号与数字之争"现在肯定已不复存在，并已有很长时间不存在了"(Newell 1983:10)。在 J·豪格兰(Haugeland 1985)或 M·博登(Boden 1977)的 AI 学科史中，罗森布拉特甚至没有被提及。[1]

但是把联结论者的失败归咎于反整体论的偏见，就过于

---

① 在心理学和神经科学中，关于神经网络的研究还勉强维持着。在布朗大学，J·A·安德森仍在为心理学中的网络模型进行辩护，尽管他不得不依靠其他研究者的补助，同时 S·格罗斯贝格研究出基本认知能力的一个漂亮的数学实现方式。安德森的立场见 Anderson(1978)。格罗斯贝格在这个萧条年代里的工作的例子见他的著作(Grossberg 1982)。科霍宁的早期工作见于《联想记忆——一种科学理论方法》(Kohonen 1977, Berlin: Springer—Verlag)。在麻省理工学院，明斯基继续就神经网络举办讲座，并指定一些研究神经网络逻辑特性的论文。但是根据帕佩尔特的看法，明斯基这样做，仅仅是因为网络具有令人感兴趣的数学特性，然而在符号系统的特性方面却证明不出任何令人感兴趣的东西。此外，许多 AI 研究者想当然地认为，既然图灵机是符号操作装置，同时图灵已经证明图灵机可以做任何计算，那么他也就证明了所有的智力能力都可通过逻辑获取。根据这一观点，整体论(以及那时的统计学)方法的合理性就需要加以确证，而符号 AI 方法则无需确证。然而这种自信是建立在把图灵机中未经解释的符号(0 和 1)和 AI 中从语义上解释的符号混为一谈的基础上的。

简单了。一些哲学假定产生了对直觉的影响，及导致了对早期符号加工成果的重要性的过高估计，还有更深入的方面。从当时来看，这个方面是，研究感知机的人即使是解决最简单的模式识别问题，如将视野各个部分中的水平线与垂直线区分开来，也不得不进行大量数学分析和计算，而符号操作方法只作出相对较小的努力，就解决了困难的认知问题，如证明逻辑定理和解决组合难题。甚至更加重要的是，根据当时达到的计算力量，神经网络研究者们看来只能做纯理论的神经科学和心理学方面的工作，而符号表述论者的简单程序却即将成为有用的东西。在对形势作出这种评价的背后，有一个假定，即思维和模式识别是两个不同的领域，而两者之中思维显得更重要。正如在后面讨论常识知识问题时将会看到的，以这种方式看待问题，既忽视了模式辨别在人类专门知识中的出色作用，也忽视了在日常现实世界的思维中被预先假定的常识性理解的背景。要考虑这一背景，就十分需要模式识别了。

这一思想把我们带回到哲学传统。支持符号信息加工的，不仅是笛卡尔和他的传人们，而且是全部西方哲学。根据海德格尔的说法，传统哲学得以确定，从一开始就是因其关注世界中的事实而“忽略”世界本身(Heidegger 1962: § 14 - 21; Dreyfus 1988)。也就是说，哲学从一开始就系统地忽视或扭曲了人类活动的日常语境[①]。此外，从苏格拉底经由柏拉图、笛卡尔、莱布尼兹和康德一直传至常规 AI 的这一哲学传统的分

① 根据海德格尔的看法，亚里士多德比其他任何哲学家更接近于理解日常活动的重要性，但是即使他也屈从于对隐含于常识之中的日常世界现象的扭曲。

支，认为对某一领域的理解在于持有这一领域的**理论**是理所当然的。理论是根据抽象的原理（包括定律、规则、程序等）对客观的、**与语境无关**的元素（简单物、原素、特征、属性、因素、数据点、线索等）之间的关系进行系统阐述的。

柏拉图认为，在诸如数学甚或伦理学这样的理论领域中，思想家们运用的是一些显式的、与语境无关的理论规则，这是他们在日常世界以外的另一种生活中学会的。这样的理论一旦被掌握，就会通过对思想家心灵的控制在这个世界中发挥作用，无论思想家是否意识到这些理论。柏拉图的论述并不适用于日常技能，而只适用于存在着先验知识的领域。然而理论在自然科学中的成功强化了这一思想：在任何有序的领域内，必然存在着某个与语境无关的元素的集合，以及这些元素之间的使这一领域有序并使人们有能力以智能方式在该领域中活动的某些抽象关系。于是，莱布尼兹大胆地对所有智能活动形式甚至日常实践总结出一套理性主义的说明：

> 在各行各业中，有关技能的最重要的观察和转化方式现在仍是空白。当我们从理论过渡到想要完成某事的实践时，我们的经历就证明了这个事实。**当然，我们也可以把这个实践写出来，因为这归根到底不过是另一个更复杂，更特殊的理论**……（黑体后加）（Leibniz 1951:48）

由于把哲学家们所创立并在自然科学中获得成功的方法用到了所有领域，符号信息加工方法便胜券在握了。既然根据这一观点任何领域必然都是可形式化的，在任何范围内实施 AI 的方法显然都是找出与语境无关的元素和原理，并把形

式的符号表述建立在这一理论分析的基础上。沿着这一脉络,T·威诺格拉德描述了他借鉴物理科学进行的 AI 研究:

> 我们关心的是形成一种形式体系,或"表述方式",用来描述……知识。我们寻求的是组建这一形式体系的"原子"和"质点",以及作用于其上的"力"(Winograd 1976:9)。

毫无疑问,关于宇宙的理论常常是逐步建立的,先形成相对简单的孤立系统的模型,再逐步使模型变得更为复杂,并把它与另一些领域中的模型结合起来。这所以可能,是因为所有现象都被推测为是帕佩尔特和明斯基所称的那些"结构原素"之间的定律式关系的后果。因为没有人主张在 AI 中作原子式的还原,看来 AI 工作者只是隐含地假定,把元素从它们的日常语境中抽象出来这一点,既然确定了哲学,并在自然科学中有效,所以在 AI 中必然也是有效的。这个假定可以很好地解释物理符号系统假设何以如此之快地变成一种新发现,以及帕佩尔特和明斯基的著作何以轻而易举地击败了感知机的整体论。

60 年代中期,本文作者之一休伯特在麻省理工学院教授哲学时,很快卷入了关于 AI 可能性的辩论。很显然,像纽厄尔、西蒙和明斯基这样的研究者们是哲学传统的继承人。但是后期维特根斯坦和早期海德格尔结论的出现,对于还原论研究纲领来说并不是一个好兆头。这两位思想家对作为符号信息加工基础的传统本身提出了质疑,他们都是整体论者,他们都被日常实践的重要性所吸引,他们也都认为人们不可能

持有关于日常世界的理论。

1953年维特根斯坦发表了他的《哲学研究》,以极其犀利的文笔批判了他自己在《论文》中的观点,这是智能史上颇具讽刺意味的一幕,因为这恰是AI采纳他所抨击的抽象原子论传统的时候。维特根斯坦在写完《论文》之后,花了数年时间从事他所谓的现象学研究(Wittgenstein 1975)——寻找他的理论所需要的原子事实和基础客体,但是一无所获。他最后终于放弃他的《论文》和所有理性主义哲学。他论证道,将日常情境分析成为事实和规则(大多数传统哲学家和AI研究者们正是认为理论必然是从这里开始的),这种做法本身只有在某个语境中和为了某个目的才有意义。因此,被选出的元素已经反映出它们因之而被创立的那些目标和目的。当我们试图找出那些终极的与语境无关、与目的无关的元素时(这正是我们为了找出本原符号供计算机之用而必须做的),实际上我们是想使我们的经验的一些方面摆脱实用主义的组织方式,而正是这种组织方式使得以智能方式用这些元素处理日常问题成为可能。

维特根斯坦在《哲学研究》中直接批判了《论文》中的逻辑原子论:

> "名字实际上表示单质,在这一思想的背后是什么呢?"——苏格拉底在《泰阿泰德篇》中说道:"如果我没有弄错的话,我听到一些人这样说:原级元素是没有定义的,而我们及其他每一样东西可以说都是由它们构成的……但是正因为由这些原级元素构成的事物本身是复杂的,所以这些元素的名字结合到一起时就成为描述语

言"。罗素的"个体"和我的"客体"(《逻辑哲学论》)都是这样的原级元素。但是构成实体的单质组分是什么呢?……绝对地说"椅子的一些单质部分"是毫无意义的(Wittgenstein 1953:21)。

在20年代,M·海德格尔就已经用类似的方式反对他的导师E·胡塞尔,胡塞尔自认为是笛卡尔传统的顶峰,因而也就是AI的师祖(Dreyfus 1982)。胡塞尔论证说,意识的活动,或意向作用(noesis),并不以其本身去掌握一个客体,相反这一活动获得意向性(定向性)所凭借的仅仅是与这一活动相关联的意向对象(noema)中的"抽象形式"或意义。[①]

这个意义,或符号表述方式,根据胡塞尔的构想,是一个使一项困难任务得以完成的复杂实体。在《纯粹现象学观念》(Husserl 1982)中,胡塞尔大胆地尝试解释意向对象是怎样完成这项任务的。提供所指的"谓词意念",像弗雷格的意义(Sinne)一样,恰恰具有辨别客体原子特性的显著特征。这些谓词被结合到复杂客体的复杂"描述"中,这与罗素描述理论的看法一样。胡塞尔在这一点上很接近康德,他认为意向对象包含着一个由严格规则组成的层级体系。由于胡塞尔把智能看成一种语境确定的、有目标导向的活动,所以任何种类的客体的心理表述都必须提供一个语境,或是一些期望的或"预先刻划"的"视界",以便使新进入的数据结构化:"一个有可能

① "正如我们所认为的那样,意识不是认识论中的一个存在着的具体事物,而是一个寄存的,以抽象形式存在的行为。"(原文为德文——译者)见胡塞尔(Husserl 1950)。有关胡塞尔认为意向对象说明精神活动意向性的证明,见H·德雷福斯的"胡塞尔的知觉意向对象"(Dreyfus 1982)。

规定(客体的)其他意识的规则可能等同于从本质上例示出预先刻划的类型的规则"(Husserl 1960:45)。意向对象必须包含描述所有特征的规则,在考察某一**类型**的客体时,无疑会预先想到这些特征,它们保持着"不容动摇的原状:只要客观性仍然被意指为**这个**和这类"(1960:53)。对于这一类型的客体的可能的而不是必要的特征,该规则也必须规定出一些预先刻划:"因此并没有一个完全确定的意念,而是始终存在着一个**空白意念的框架**……"(1960:51)。

1973年M·明斯基为表述日常知识提出了一个新的、与胡塞尔所说的极其相似的数据结构:

> 一个**框架**就是一个用来表述陈规情境如待在某一种居室里或是去参加儿童生日聚会的数据结构……
>
> 我们可以把框架看成由节点和关系构成的网络。框架的顶层是固定的,所表述的事情对所设情境来说总是真的。较低层次上有许多**终端**——一些必须用特殊例子或数据填充的槽。每一终端都能够详细说明它的作业必须满足的条件……
>
> 该理论的现象学力度,在很大程度上取决于期望和其他各种推测所包含的内容。**框架的终端在一般情况下已由"缺省"分配所填充**(Minsky 1981:96)。

在明斯基的框架模型中,"顶层"是由按照胡塞尔的说法在表述中保持"不容动摇的原状"的东西发展而来的,而胡塞尔的预先刻划则已变成"缺省分配"——可按正常方式预期的新增特征。其结果是,AI的技术向前迈进了一步,从被动的

信息加工模型,发展为试图考虑认知者与世界间相互作用的模型。这样,AI的任务就与先验现象学的任务一致了。这两者都必须尝试在日常领域中找出由一组本原谓词和它们的形式关系所构成的框架。

作为对胡塞尔的回应,海德格尔先于维特根斯坦完成了对日常世界和像椅子、榔头这样的日常客体的现象学描述。和维特根斯坦一样,他发现日常世界不能用一组语境无关的元素来表述。海德格尔指出,除了把事物作为由一组谓词定义的客体与之发生联系的方式之外,还存在着另一些与事物"相遇"的方式,正是他通过这一点迫使胡塞尔明确正视这一问题。海德格尔说,当我们使用一件像榔头这样的工具时,我们是在一个由(不需要表述为一组事实的)工具、目的和人的角色通过社会方式组织起来的连接关系的语境中,使一种(不需要在心灵中表述的)技能变为现实。这个语境,或世界,以及我们在其中应付自如的日常方式,即海德格尔所谓的"周全性",并不是我们**想出**的什么东西,只是我们社会活动的一部分,它形成了我们的**存在**方式。海德格尔总结道:

> 在关系系统的意义上,语境……可以被形式地对待。但是……这些"关系"和"被关系者"的现象内容使它们对任何一种数学机能化作出抵制;它们也不仅仅是事先被安置在一个"思维行动"中的被思考的某件事情。它们更加是这样一些关系,其中已经存在着有关联的周全性本身(Heidegger 1962:121 - 1)。

这确定了胡塞尔和AI一方与海德格尔和后期维特根斯

坦一方之间的种种方式的分裂。关键问题成了:“理性主义哲学家们所坚持的那种关于日常世界的理论,有可能存在吗?”或毋宁说常识背景是技能、实践、辨别力等等的结合吗?技能、实践、辨别力等不是意向状态,更不用说它们是没有根据元素和规则解释的任何表述内容的。

胡塞尔采取了一个很快在AI圈中变得为大家所熟悉的手法,试图绕过海德格尔提出的问题。胡塞尔认为,世界、重要背景、日常语境,只不过是一个非常复杂的系统,该系统是由一些与一个复杂的信念系统相联系的事实组成的,因为这些信念具有真值条件,所以他称之为有效性。他认为,原则上讲,一个人可以将自己在世界中的存在悬置起来,而完成对人类信念系统的独立描述。这样,人就能完成那个隐含在从苏格拉底以来的哲学中的任务:人们能使那些作为全部智能行为的基础的信念和原理变得明确。正如胡塞尔所指出的:

> 即使我们总是同时意识到但暂时无关紧要因而一直完全不被注意的那种……背景,仍然根据它隐含的有效性在发挥作用(Husserl 1970:149)。

因为胡塞尔坚信可以使共享的背景像一个信念体系那样外显,所以他先于他的时代提出了AI可能性的问题。在讨论了一个形式公理系统可以描述经验的可能性,同时指出这样的公理和原素系统——至少像我们在几何学中所知道的——不能描述像“扇贝形”和“透镜形”这样的日常形状之后,胡塞尔留下这个关于这些日常概念是否仍能被形式化的问题没有解决。(这与提出有关人们是否能将常识物理学形式化的AI

问题而留下它没有解决的情况是一样的。)在承继莱布尼兹关于建立全部经验的数学的梦想时,胡塞尔又写道:

> 迫切的问题是……是否不可能存在……一个用纯粹而严密的理念代替直觉资料的、同时可以……作为……经验数学基本手段的理想化过程(Husserl 1952:v.134)。

但是正如海德格尔所预见的那样,对日常生活作出完全理论叙述的任务其实比最初预想的要困难得多。胡塞尔的课题遇到很大的麻烦,一些迹象表明明斯基的课题也遇上了麻烦。在尝试说清楚主体对日常客体的表述的各个成分的二十五年间,胡塞尔发现,他不得不把越来越多的主体对日常世界的常识性理解包容进来:

> 毫无疑问,即使那些当我们把单一类型的客体当做有限的线索时自行呈现出来的任务,也被证明是极其复杂的,当我们向纵深发展时,这些任务总是导致广泛的学科。例如……空间客体(更不用说自然界)本身的情况就是这样,心理物理存在和人类本身以及文化本身的情况也是如此(Husserl 1960: 54-5)。

他谈到了意向对象的"臃肿的具体形式"(Husserl 1969:244)和它的"惊人的复杂性"(1969:246),在75岁的时候,他沮丧地得出结论:他永远是一个初学者,而现象学是一个"永无止境的任务"(1970:291)。

在明斯基的"表述知识的框架"一文中,有迹象表明他已

开始从事于那个最终将胡塞尔击垮的同一个“永无止境的任务”:

> 仅是建构一个知识基础,就成为智能研究的重大问题……关于常识性知识的内容和结构,我们还是知道得太少了。“极小”常识系统必须“知道”有关因果、时间、目的、地点、过程和知识类型……的某些情况。在这一领域中,我们需要花力气做严格的认识论研究(Minsky 1981:124)。

学习当代哲学的学生对明斯基的天真和信心会感到吃惊。胡塞尔的现象学正是这样的研究工作。的确,从苏格拉底经莱布尼兹到早期维特根斯坦,哲学家们两千年来在这一领域中进行了严格的认识论研究,但没有取得显著的成就。

鉴于维特根斯坦的转向,和海德格尔对胡塞尔的强有力的批评,本文作者之一——休伯特——预见到符号信息加工的麻烦。正如纽厄尔在他的 AI 史中所说的,这一警告被忽视了:

> 德雷福斯主要的、智力方面的反对意见……是:将人类活动的语境分析为离散的元素是注定要失败的。这一反对意见建立在现象学哲学的基础上。遗憾的是,就 AI 而言,这意见像是不存在一样。接踵而来的对德雷福斯文章所作的回答、驳斥和分析,根本没有针对这一论点——如果它要走向前台的话,它的确还是个新论点(Newell 1983:222-3)。

的确，这个麻烦没有多久就走向前台了，就像对传统哲学的报复一样，日常世界又报复了 AI。正如我们看到的那样，由纽厄尔和西蒙首创的研究纲领已经经历了三个十年阶段，从1955 至 1965 年，表述和搜索这两个研究主题，在当时称为“认知模拟”的领域占据了主导地位。例如，他两人指出计算机怎样能用被称为手段目的分析的一般启发式搜索原理来解决一类问题，即采用任何一种有效的运作方式，以减小当前状况描述与目标描述之间的距离。然后他们把这种启发式技术提取出来，并纳入他们的“一般问题求解程序”(GPS)。

第二个阶段(1965—1975)由麻省理工学院的 M·明斯基和S·帕佩尔特领导，主要涉及要表述的是何种事实和规则的问题。其思想是，在称为“微世界”的孤立领域中，形成系统地处理知识的方法。1970 年前后在麻省理工学院编写的著名程序有：T·威诺格拉德的能接受在自然语言子集合中给出的有关简化“积木世界”的指令的 SHRDLU，T·埃文的类比问题程序，D·沃尔兹的场景分析程序和 P·温斯顿的能根据例子学习概念的程序。

人们希望这些限定的、孤立的微世界能够逐步变得更接近现实，并且联合起来，以便通往对现实世界的理解。但是研究者们混淆了根据海德格尔的观点我们应区分为“全域”和“世界”的两种领域。一组相互联系的事实可以构成一个**全域**，像物理全域，但是不能构成一个**世界**。后者如商业界、戏剧界或物理学家界，是由客体、目标、技能和实践组成的团体，在它的基础上，人类活动才有意义或是讲得通。为了弄清这一区别，可以将**没有意义的**物理全域和**有意义的**物理学科界

加以对比。物理学界、商业界和戏剧界只有靠着人类共同关切的背景才有意义。它们是全体人类共享的一个常识世界中的局部精制品。也就是说，子世界不像可孤立的物理系统那样，与那个由它们**组成**的更大的系统发生联系，而是它们**预设**整体中的局部的人为作品。微世界并不是世界，而是孤立的、无意义的领域，并且人们渐渐看清了，没有一种可以使它们联合起来并扩展到抵达日常生活世界的方法。

第三个阶段约从1975年到现在，在这期间，AI一直在同后来人们所说的常识知识问题进行较量。知识表述一直是AI工作的中心问题，但前两个阶段——认知模拟和微世界——的特点是弄清用尽可能少的知识能做多少事情，试图避免常识知识问题。然而到70年代中期，已无法再回避这个问题。各种数据结构，如明斯基的框架和香克的脚本，都已尝试过而没有成功。常识知识问题甚至使AI无法开始实现西蒙在二十年前的预言："二十年之内，机器将能做到人所能做的一切"(Simon 1965:96)。

的确，在过去的十年间，常识知识问题成为理论AI所有进步的障碍。威诺格拉德是最早认清SHRDLU程序的局限性和所有脚本及框架试图扩展微世界方法的局限性的人之一。对AI"失去信心"之后，现在他在斯坦福的计算机科学课程中讲授海德格尔，并指出"困难在于把那种确定哪些脚本、目标和策略是相关的以及它们怎样相互作用的常识背景形式化。"(Winograd 1984:142)。

使AI困于这一僵局之中的原因，是坚信常识知识问题必定是可以解决的，因为人类显然已经解决了这一问题。但是人类很可能根本不是按通常方式使用常识性知识的。正如海

德格尔和维特根斯坦所指出的，与常识性理解相当的，很可能是日常技能。所谓"技能"，并不是指过程的规则，而是指在众多的特定场合知道该做什么。[1] 例如，人们曾经发现常识物理学原来是极难用一组事实和规则详加说明的。如果试图这样做，那么，或者需要更多的常识去理解所发现的事实和规则，或者构造出一些复杂得看起来决不可能存在于儿童的心灵中的公式。

从事理论物理学工作，也需要一些背景技巧，它们也许是不可形式化的，但是该领域本身可以用不参考这些背景技巧的抽象定律来描述。AI 研究者们错误地得出结论：常识物理学也必然可以表示为一组抽象原理。然而寻找常识物理学理论的问题恰恰可能是不可解的，因为这是一个没有理论结构的领域。一个孩子在每天玩弄所有类型的液体和固体数年后，有可能直接学会辨别各种固体、液体等等的原型情况，并且学会在典型环境中对它们的典型行为作出典型而熟练的响应。同样的情况也很可能出现在社会现实中。如果背景理解的确是一种技巧，而技巧又是以全体模式而不是以规则为基础的，那么我们就可以预期符号表述方式不能获取人类的常识性理解。

由于走进死胡同，经典的、基于符号的 AI 显得越来越像 I·拉卡托斯(Lakatos 1978)所说的那种退化中的研究纲领的绝妙例子。正如我们已经看到的那样，AI 以纽厄尔和西蒙在兰德的工作兴盛地开始，到 60 年代后期成为一个蒸蒸日上的研

[1] 关于技能的这一说明，作者在另一文章(Dreyfus and Dreyfus 1986)中有详细而雄辩的阐述。

究纲领。明斯基曾预言,“只需一代人的时间,创造‘人工智能’的问题就可以基本解决”(Minsky 1977:2)。然后很突然地,这个领域就碰上了未曾料到的困难。把常识阐述成理论,原来比人们设想的困难得多。它并不像明斯基所希望的那样,仅仅是一个为成千上万的事实编写目录的问题。常识知识问题成为被关注的中心。五年后,明斯基的情绪完全改变了。他对一个记者说:“AI 问题是科学曾从事研究的最困难的问题之一”(Kolata 1982: 1237)。

理性主义传统终于被置于经验检验之下,而它失败了。为日常的常识世界建立一个形式的原子论理论的思想,以及用符号操作器来表述该理论的思想,恰恰遇到了海德格尔和维特根斯坦已经发现的困难。F·罗森布拉特直觉地感到,使世界形式化,从而形式地说明智能行为,面临着无法逾越的困难,这种直觉已得到证实。他那个受压制的研究纲领(用计算机例示理想化大脑的整体论模型),从未真正被驳倒,现在又变成一个有生命力的选择。

在对 AI 历史的新闻报道中,一些匿名攻击者把罗森布拉特诬蔑为江湖郎中:

> 现今的研究者们都记得,罗森布拉特习惯于对他的机器的性能不断作出夸张的说明。一位科学家说,“他梦想成为一个新闻宣传者,实际上是一名游医。根据他的说法,感知机能够做一些十分奇特的事情。也许感知机能这样做,但是从罗森布拉特的工作中,这一点是无法得到证实的”(McCorduck 1979:87)。

其实，比起西蒙和明斯基对他们的符号程序的认识来，他对各种类型的感知机的能力和局限性的认识要清楚得多。[①] 现在他的名誉正在恢复。D·鲁梅哈特、G·欣顿和J·麦克莱兰对他的开拓性工作重新表示赞赏：

---

① 罗森布拉特《神经动力学原理》一书中某些典型观点的摘录（Rosenblatt 1962a）：在学习实验中，典型的做法是向感知机展示一系列模式，这些模式含有要加以辨识的每一类型或类别的典型事例，同时，根据某个供记忆修正用的规则，响应的恰当选择得到"强化"。然后感知机得到一个试验刺激，对这类刺激作出恰当响应的概率被确定……如果试验刺激激活一组感觉元素，而这组元素完全不同于以前呈现的同类刺激所激活的那些元素，那么，这个实验就是一个"纯概括"试验。最简单的感知机……没有纯概括能力，但是可以证明，它在辨别实验中表现得相当不错，特别是当试验刺激与以前经历过的模式之一几乎等同的时候（p68）……迄今为止考虑过的感知机表明，在图形探测能力方面，以及在格式塔组织倾向方面，它们与人类受试者的相似之处很少（p71）……对初等形式序列的识别，完全在组织得当的感知机的能力范围之内，但是图形组织和分割问题产生出许多问题，其严重程度和静态模式感知中的情况一样（p72）……在简单感知中，模式先于"关系"被识别，的确，有些抽象关系，诸如"A在B上方"或"三角形在圆内部"，从未真正得到抽象，而只能借助于详尽无遗的死记硬背的学习过程来获取，在这过程中，保持该关系的每一种情况被逐一地教给感知机（p73）……由少于三层的信号传递单元构成的网络，或是仅由按串行方式联结的线性单元构成的网络，不可能学会辨别各向同性环境中的模式类别（在这种环境中，任何模式都可以出现在所有可能的视网膜位置上，而不产生边界影响）（p575）……在前面的几章中，已经提出了若干在研究之中的模型，它们有可能学习序列程序，将言语分析成音位，还能够通过简单的感觉对象来学习名词和动词的实际"意义"。这些系统代表着迄今为止考虑过的感知机中抽象行为的上限。它们也有缺陷，因为它们缺乏令人满意的"暂时记忆"，没有能力感知简单样式的抽象拓扑关系，并且除了在特殊条件下以外没有能力分离出有意义的图形实体或客体（p577）……从本书提到的种种感知机来看，最有可能实现的应用方面有：特征识别和"阅读机"，对（单词清晰可辨的）言语的识别，以及十分有限的图片识别能力，或是对简单背景下的对象的识别。较宽意义上的"感知"很有可能是我们当前模型的后继者们所能掌握的，但是必须在获取相当大量的基础知识之后，才能制定出一个充分精巧的设计方案，使得感知机在正常环境条件下可与人竞争（p583）。

当时对罗森布拉特的工作争议很大，同时他提出的特定模型也未能达到他所寄予的全部希望。但是他把人类信息加工系统看作动态的、相互作用的、自组织的系统，这种看法却是PDP方法的核心(Rumelhart and McClelland 1986:i.45)。

感知机的研究……清楚地预见到许多对今天有用的结果。明斯基和帕佩尔特对感知机的批评，产生了广泛误解，损害了感知机的信誉，而那本书只是说明了感知机式机构中力量最有限的一类的局限性，对更有力的多层模型则只字未提(1986:ii.535)。

受挫的AI研究者们，厌倦于依附那个被J·莱特温在80年代初描绘成"仅有一根漂浮着的稻草"的研究纲领，纷纷拥向新的范式。鲁梅哈特和麦克莱兰的《并行分布式处理》一书刚问世就售出6000册，到现在已经印刷了3万册。正如P·斯莫伦斯基所指出的：

近五年来，有关认知模型的联结论方法已经从只有少数几个忠实信徒的默默无闻的教派成长为一个如此强劲的运动，以致最近几次认知科学学会会议，开始给人以像联结论者在举行鼓动大会的印象(待出版)。

如果多层网络成功地实现了它们的期许，研究者们就不得不放弃笛卡尔、胡塞尔和早期维特根斯坦认为产生智能行为的唯一方式是在心灵中用形式理论来反映世界的信念。更坏的是，人们也许不得不放弃对哲学本源的更为基本的直觉，

即认为必定存在着一个有关现实每一方面的理论，也就是说，必定存在着一些元素和原理，借助它们，人们就能对任何领域的智力能力作出说明。神经网络有可能表明，海德格尔、后期维特根斯坦和罗森布拉特认为我们在世界上的智能行为并不需要有关这个世界的理论的想法是正确的。如果理论不是解释智能行为所**必要的**，我们就只好准备提出这样的问题：在日常领域中，这样的理论解释是否还是**可能的**。

在符号操作 AI 的影响下，神经网络模型的建立者们正在尽力而为，一旦他们的网络被训练得能完成一个任务，就试图找出由个别节点和节点组所表述的特征。其结果至今还是可疑的。我们来看看欣顿利用分布式表述建立的概念学习网络(Hinton 1986)。在一个人类借助特征实施概念化的领域中，即使没有把人类所用特征给予网络，网络也能被训练得对该领域中的关系进行编码。欣顿得出了一些情况的例子，在这些情况下，对受训网络中某些节点的解释可以达到符合人类分辨特征的程度，尽管这些节点只是大体上符合那些特征。然而大多数节点根本不可能在语义上得到解释。符号表述方式中所用的特征可能出现，也可能不出现。然而在网络中，虽然当某一特征出现在这一领域中时某些节点显得更为活跃，但是总的活动性不仅随着这一特征的出现与否而发生变化，而且也受到其他特征的出现与否的影响。

欣顿选择了一个领域——家庭关系，该领域是由人类严格按照人类通常注意到的那些特征如代别和国籍构成的。然后，欣顿对这样一些事例进行了分析：在这些事例中，从某种随机初始联结强度开始，经过学习，某些节点就可以被解释为代表了这些特征。可是使用欣顿模型的演算表明，甚至在完

全没有明显使用这些日常特征的情况下，他的网络似乎也学会了以其联想取代某些随机的初始联结强度。

从十分有限的意义上来看，任何训练有素的多层网络都可以就特征作出解释——不是日常的特征，而是那种我们将称为高度抽象特征的东西。我们来看一个简单情况，由一些被正反馈而不是横向或负反馈联结激活的二进制单元组成的多层机。为了由已经学会某些联想的网络来构成这样的说明，输入节点上面一层中的每一节点可以依据与它的联结方式被解释为在对一组特定输入模式中出现的一个模式进行探测。（有些模式将成为训练中使用的模式，有些模式则永远不被使用。）如果要把一个构想的名字（几乎可以肯定我们的词汇表中没有它的名字）给予一个特定节点探测到的一组输入模式，那么这个节点就可以被解释为是在探测以此命名的高度抽象的特征。这样，输入层的上面一层中的每一节点，就能够被刻划成特征探测器。类似地，这些节点上面一层中的每一节点，可以被解释为在探测更高阶的特征，这特征被定义为在第一层次的特征探测器中存在着一组经过详细说明的模式之一。还可循着层级体系向上继续进行。

智能被定义为有关适合于一领域的一组特定的联系的知识，根据一个技能领域中许多高度抽象的特征之间的关系总是可以对智能加以说明，然而这一事实并没有维护理性主义的直觉，即认为这些解释特征必须获取该领域的基本结构，以便能在它们的基础上建立理论。如果再多把一个输入输出对的联系（这里，训练之前的输入所产生的输出，不同于要学习的那个输出）教给网络，至少有些节点的解释就必须改变。所

以在最后一个训练的例子之前，由某些节点分辨出的特征原来并不是该领域的不变的结构特征。

人们一旦放弃了经典 AI 的哲学方法，并接受了神经网络建模的非理论主张，那就留下这样一个问题：能够预期这样的网络获得多少日常智能呢？正如鲁梅哈特已经注意到的，经典 AI 研究者很快就指出，迄今为止，神经建模者已在按步长进行问题求解方面遇上了困难。联结论者的答复是：他们有信心，他们迟早会解决这个问题。然而这一答复太容易使人想起 60 年代符号操作者们对于批评他们的程序在模式感知方面的欠缺所作的那种答复了。以前的斗争还在理智主义者与格式塔主义者之间继续着，理智主义者认为，由于他们能够进行语境无关的逻辑活动，所以具备处理日常认知的能力，但是在理解知觉方面有些欠缺，格式塔主义者则有一些关于知觉的初步说明，但不能对日常认知方式作出说明。① 用了右大脑和左大脑的比喻，人们可能会认为，大脑或心灵也许会运用每一个适当的策略。这样，就会产生怎样使这些策略结合起来的问题。人们不能只是来回转换，因为据海德格尔和格式塔主义者的看法，在相关性的确定中，甚至在日常逻辑和问题求解中，起关键作用的是有实际价值的背景，而任何领域甚至逻辑方面的专家，都是利用它们在功能上的相似性来掌握运算的。

考虑将这两种方法结合起来还为时过早，因为到目前为止，没有一种方法已经作出了足够的成绩，使它具有坚实的基

---

① 最近一个很有影响的关于知觉的说明否定了对心理表述的需要，见吉布森(Gibson 1979)。吉布森和罗森布拉特 1955 年为美国空军合作了一篇研究论文，见吉布森、奥卢姆和罗森布拉特(Gibson, Olum and Rosenblatt 1955)。

础。神经网络建模也许只是获得了一个应得的失败机遇，就像符号方法经历过的那样。

当然，我们不应忘记每个研究纲领的奋斗目标有着重大差别。物理符号系统方法之所以会失败，看来是因为关于每一领域都必定有一个理论的假定完全是错误的。然而神经网络建模并不受这一假定或任何别的哲学假定的约束。不过，建立一个与人类大脑进化而成的网络充分相像的相互作用网络，也许过于困难了。的确，十五年来常识知识问题已经成为符号表述技术进步的障碍，但是它也许在昭示着神经网络的来临，尽管研究者们也许并未意识到这一点。所有神经网络建模者都认为，一个网络要具有智能，它必须具备概括能力，就是说，如果充分给出一些与一个特殊输出相联系的输入的例子，它就应当把更多同样类型的输入与同样的输出联系起来。然而这就出现了一个问题：什么算作相同类型呢？对于进行合理概括所需要的类型，网络设计者的思想中有一个专门的定义，如果这个网络对这一类型的其他例子作出了概括，就认为它是成功的。但是当这个网络产生了一个意外的联想系时，我们能说它没有作出概括吗？我们同样有理由说，这个网络一直都在按照当前类型的不同定义进行活动，并且这种不同刚刚被揭示了出来。（在作智能试验时发现的"将这一序列继续下去"的问题，实际上都具有不止一种可能的答案，但是大多数人都意识到什么是简单和合理的，因而是可接受的。）

神经网络建模者试图避免这种歧义，而使网络产生"合理的"概括，方法是只考虑一族事先规定为许可的概括，即可以算作可接受概括（前提空间）的许可变换。于是这些建模者们

试图为他们的网络设计一种构造体系，使这些网络只用它们存在于前提空间中的方式将输入变换为输出。这样，概括就只能按照设计者的条款进行。要唯一地确认前提空间中的恰当成员，少数几个例子是不够的，然而有了足够的例子之后，就可以用仅有的一个前提说明所有的例子。于是，网络就将学会恰当的概括原理。也就是说，以后的所有输入都将产生出那种在设计者看来是恰当输出的东西。

这里有一个问题，根据这种网络的构造体系，设计者们已经规定了某些可能的概括是决不会被发现的。对玩具问题来说，这一切都无所谓，因为在这些问题中，不存在由什么构成一个合理概括的问题，但是在现实世界的情境中，人类智能基本上在于对语境以恰当方式进行概括。如果设计者把网络限制于预先定义的一类恰当的响应，这个网络就会表现出设计者就这一语境植入其中的智能，而不会使它像真正的人类智能那样具有能适应其他语境的常识。

如果网络和我们一样要具有恰当概括的意识，它也许必须具有和人类大脑一样的尺寸、构造和初始联结构形。如果它要从自己的“经验”学会作出人类式的联系，而不是被教会作出已经由训练者规定好的联系，它就必须也具有和我们一样的关于输出恰当性的意识，而这就意味着它必须具有和我们一样的需求、欲望和情感，而且必须有一个人类式的躯体，该躯体能做恰当的物理运动、具有种种能力，也易受伤害。

如果海德格尔和维特根斯坦是正确的，人类的整体性将比神经网络大得多。智能必须受到有机体中目的和有机体从当前文化中获得的目标的促动。如果分析的最小单元就是与整个文化世界相啮合的整个有机体的最小单元，那么神经网

络以及以符号编程的计算机，就还有漫长的道路要走。

参考书目

Anderson, J. A. (1978). 'Neural Models with Cognitive Implications.' In D. LaBerse and S. J. Samuels (eds.), *Basic Processing in Reading*, Hillsdale, NJ: Erlbaum.

Boden, M. (1977). *Artificial Intelligence and Natural Man*. New York: Basic Books.

Dreyfus, H. (1979). *What Computers Can't Do*, 2nd edn. New York: Harper & Row.

—— (1988). *Being-in-the-World: A Commentary on Division I of 'Being and Time'*. Cambridge, Mass.: MIT Press.

—— (ed.) (1982). *Husserl, Intentionality and Cognitive Science*. Cambridge, Mass.: MIT Press.

—— and Dreyfus, S. (1986). *Mind Over Machine*. New York: Macmillan.

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton-Mifflin.

—— Olum, P., and Rosenblatt, F. (1955). 'Parallax and Perspective During Aircraft Landing.' *American Journal of Psychology* 68: 372–85.

Grossberg, S. (1982). *Studies of Mind and Brain: Neural Principles of Learning, Perception, Development, Cognition and Motor Control*. Boston: Reidel Press.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea*. Cambridge, Mass.; MIT Press.

Hebb, D. O. (1949). *The Organization of Behavior*. New York: Wiley.

Heidegger, M. (1962). *Being and Time*. New York: Harper & Row.

Hinton, G. (1986). 'Learning Distributed Representations of Concepts.' In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*. Amherst, Mass.: Cognitive Science Society.

Hobbes, T. (1958). *Leviathan*. New York: Library of Liberal Arts.

Husserl, E. (1950). *Ideen Zu Einer Reinen Phänomenologie und Phänomenologischen Philosophie*. The Hague: Nijhoff.

—— (1952). *Ideen Zu Einer Reinen Phänomenologie und Phenomenologischen Philosophie*, bk. 3 in vol. 5, *Husserliana*. The Hague: Nijhoff.

—— (1960). *Cartesian Meditations*, trans. D. Cairns. The Hague: Nijhoff.

—— (1969). *Formal and Transcendental Logic*, trans. D. Cairns. The Hague: Nijhoff.

—— (1970). *Crisis of European Sciences and Transcendental Phenomenology*, trans. D. Carr. Evanston: Northwestern University Press.

—— (1982). *Ideas Pertaining to a Pure Phenomenology and to a Phenomenological Philosophy*, trans. F. Kersten. The Hague: Nijhoff.

Kohonen, T. (1977). *Associative Memory: A System-Theoretical Approach*. Berlin: Springer-Verlag.

Kolata, G. (1982). 'How Can Computers Get Common Sense?' *Science* 217 (24 Sept.): 1237.

Lakatos, I. (1978). *Philosophical Papers*, ed. J. Worrall. Cambridge: Cambridge University Press.
Leibniz, G. (1951). *Selections*, ed. P. Wiener. New York: Scribner.
Lighthill, Sir James (1973). 'Artificial Intelligence: A General Survey.' In *Artificial Intelligence: A Paper Symposium*. London: Science Research Council.
McCorduck, P. (1979). *Machines Who Think*. San Francisco: W. H. Freeman.
Minsky, M. (1977). *Computation: Finite and Infinite Machines*. New York: Prentice-Hall.
—— (1981). 'A Framework for Representing Knowledge.' In J. Haugeland (ed.), *Mind Design*, pp. 95–128. Cambridge, Mass.: MIT Press.
—— and Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. Cambridge, Mass.: MIT Press.
Newell, A. (1983). 'Intellectual Issues in the History of Artificial Intelligence.' In F. Machlup and U. Mansfield (eds.), *The Study of Information: Interdisciplinary Messages*, pp. 196–227. New York: Wiley.
—— and Simon, H. (1958). 'Heuristic Problem Solving: The Next Advance in Operations Research.' *Operations Research* 6 (Jan.–Feb.): 6.
—— (1981). 'Computer Science as Empirical Inquiry: Symbols and Search.' In J. Haugeland (ed.), *Mind Design*, pp. 35–66. Cambridge, Mass.: MIT Press.
Rosenblatt, F. (1958). *Mechanisation of Thought Processes: Proceedings of a Symposium Held at the National Physical Laboratory*, Vol. 1. London: HMS Office.
—— (1962*a*). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Washington, DC: Spartan Books.
—— (1962*b*). 'Strategic Approaches to the Study of Brain Models.' In H. von Foerster (ed.), *Principles of Self-Organization*, Elmsford, NY: Pergamon Press.
Rumelhart, D. E., and McClelland, J. L. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 2 vols. Cambridge, Mass.: MIT Press.
—— and Norman, D. A (1981). 'A Comparison of Models.' In G. Hinton and J. Anderson (eds.), *Parallel Models of Associative Memory*, pp. 3–6. Hillsdale, NJ: Erlbaum.
Simon, H. (1965). *The Shape of Automation for Men and Management*. New York: Harper & Row.
Smolensky, P. [1988]. 'On the Proper Treatment of Connectionism.' *Behavioral and Brain Sciences* [11: 1–74].
Winograd, T. (1976). 'Artificial Intelligence and Language Comprehension.' In *Artificial Intelligence and Language Comprehension*, Washington, DC: National Institute of Education.
—— (1984). 'Computer Software for Working with Language. *Scientific American* (Sept.): 142ff.
Wittgenstein, L. (1948). *Last Writings on the Philosophy of Psychology*, Vol. 1, trans. corrected. Chicago: University of Chicago Press, 1982.
—— (1953). *Philosophical Investigations*. Oxford: Basil Blackwell.
—— (1960). *Tractatus Logico-Philosophicus*. London: Routledge & Kegan Paul.
—— (1975). *Philosophical Remarks*. Chicago: University of Chicago.

14 认知神经生物学中的某些简化策略

P·M·丘奇兰*

1. 引　言

这篇文章旨在使哲学家们了解神经科学中近来正在探究的一个引人入胜的理论方法,这是一个关于表述和计算的方法。这个方法之所以引人入胜,理由至少有三。第一,对于脑怎样可能对它所处世界的诸多方面作出**表述**这一问题,该方法提供了一个极其一般性的答案。我将在本文稍后探究这一答案,将它应用于哲学家们熟悉的例子:存在于人的多种感觉直觉中的各种主观感觉特性的例子。那时我们将看到对常见的感觉特性作纯粹神经生物学简化的概况。然而,这种应用只不过是我想要说明的诸多方面之一。一个重要结果是,形形色色的表述方式,那些在常识看来特点全然不同的例子,其基础原来是完全相同的。因而,这种方法在多样性中发现了统一性。

第二个引人入胜的方面与**计算**问题有关。这里概述的表述方式,极其适合于功能强大的计算形式,即一种十分适宜于

解答多种问题的形式。这些问题中有一个哲学家们不太熟悉的问题:感觉运动协调问题。不论这个问题是多么不为人熟悉,它对认知理论来说却是极其重要的,因为根据当前经验来支配恰当的行为,正是智能的原始起点。既然感觉运动协调是任何动物都必须解决的最基本的问题,那么同时又对表述作出一般性说明的解答手段,当然会引起我们的好奇心。

第三,这种方法的引人入胜之处还表现在:它掌握了脑的**微观物理组织**的奥秘,并解答了它的特定组织如何**实现**整个脑所表现出的表述活动和计算活动的问题。同时,这种方法也获得经验的支持,因为至少有两种重要方式,对被提出的表述和计算实现了物理上的抽象处理,其中每一种都与在整个经验的脑中表现得十分突出的实际神经结构具有易使人产生联想的相似性:大脑皮质的分层组织和小脑皮质的稠密正交基质(matrix)。

总的说来,对于用神经科学解释各种常见的认知现象,这种方法构造出一个大胆的简化策略。当然,该策略得当与否,在心灵哲学内部仍是一个争论激烈的问题。历史地看,这一争论已经平息下来,这是由于在神经生物学中缺乏完全令人信服的一般性**理论**,即实际上有可能对某一常见类别的认知现象作出神经生物学简化的理论。如果在其他学科的文献中确实存在着这样的理论,至少作者没有设法使其参与哲学争论。在缺乏相关理论的情况下,反对简化论的论据得以被提

---

* P·M·丘奇兰所著"认知神经生物学中的某些简化策略"来自《心灵》(Mind) XCV, no.379,1986 年 7 月:279—309。牛津大学出版社允许重印,并作了几处修改。

丘奇兰(Paul Churchland),加利福尼亚大学(圣迭戈)认知科学系哲学教授。

出，并确实一再被提出，这时只要举出我们认知活动的某些方面，并提出这种修辞学上的问题："究竟怎样才能通过可能对神经元基本要素作出的任何陈述，来说明'**这**'，甚或为'**这**'编址？"

这种修辞学的问题不公正地利用了我们想象力的软弱性，因为即使对这种现象作远非恰当的回答，也没有理由指望我们能按要求把它构想出来。但是凑巧的是，一些潜在恰当的回答确实出现在近期认知心理学和认知神经生物学的研究中。我相信，它们的存在必然很快把我们的注意力从简化是否可能这一抽象问题，转移到一个具体问题上：在各种可供选择的神经生物学理论中，哪一个真正给出了正确的简化，同时也转移到它对人类整个自我概念的长远影响上。

正如我们要解释的那样，这一基本思想是：脑根据在适当**状态空间中的位置**，对现实的各个方面作出表述；同时脑通过从一个状态空间到另一个状态空间的一般**坐标变换**根据这种表述来完成计算。这些观念看起来也许是神秘的和难以接受的，但下面的几个图会迅速解除它们的神秘性。这一理论完全是可以理解的，在其最简形式中，的确可以从**直观上**作出理解，甚至不懂数学的读者也完全可以理解。

我最初对这种理论方法有所了解，是由于阅读了神经科学家 A·派利欧尼斯和 R·利纳斯（A. Pellionisz and R. Llinas 1979, 1982, 1984, 1985）富有刺激性的文章。他们的讲解比本文所作的概述要一般化得多，透彻得多。但为了讲解的方便，对他们的开创性工作的讨论将推迟到本文的后面几节。我希望开始时使事情尽可能简单些。

那么我们就提出三个显然不同的难题作为我们讨论的开

始：

1. 脑怎样**表述**世界以及脑怎样根据这些表述来执行**计算**的奥秘，

2. **感觉运动协调**的奥秘，以及

3. 脑的**微观物理组织**的奥秘。

特别令人鼓舞的是，这些问题看来有可能同时得到解答。我们从对付第三个奥秘即微观结构奥秘开始。

2. 分层大脑皮质、纵向联系和拓扑形态映射图

脑的大脑半球表面包含一个薄层，标准说法是“灰质”，这是大脑半球大部分神经细胞胞体的所在地（见图 14－1a）。剩下的“白质”主要由一些长轴突组成，它们从灰质层细胞伸向脑的其他部分。如果仔细观察这个凹凸不平的皮层的内部结构，可发现它进一步细分为数层（见图 14－1b）。人类大脑皮质分 6 层。其他动物表现出不同的层数，但分层模式是标准的。

根据每个亚层的细胞类型和集中度，以及每个亚层内的大量层面纤维即“水平”突起，可以将这些细分的层区分开来。此外，这些不同的层还可通过它们的特异性传入和传出联系进一步明确区分。表面数层只接受来自感觉周围、或大脑皮质其他部分、或脑的其他部分的某种信息输入。而深层无例外地都是信息输出层。

最后，通过大量的可以使层与层之间作交流的纵向细胞，

把这大脑皮质各层有条理地联系在一起,如同钉子穿入胶合板那样。这些纵向神经细胞从在上的表面传入层到深部的传出层,“自上而下”地引导神经元的活动。

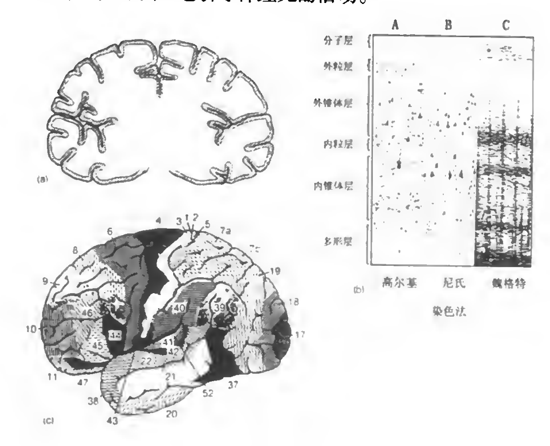


图 14-1

(a) 大脑半球剖面图,示表面灰质层:大脑皮质。(b) 以三种不同染色法显示的大脑皮层内部分层结构。(c) 布罗德曼区域。

如果现在我们离开显微镜沿周边作透视,从外面观察大脑表面的皮质层,我们就会发现它是由许多较小区域组成的拼装物(见图 14-1c)。根据它们分层细胞构筑的差异,可以在一定程度上将这些区域区分开来。这些区域最初就是基于这一判据按照它们的发现者命名为**布罗德曼区域**的。这些区域,或它们的分区,还有进一步的意义,因为其中若干个区域直接构成了感觉或运动终末的、或是脑的其他某个区域的某一方面的**拓扑形态映射图**。例如,脑后部视皮质的给定层中

细胞之间的相互位置关系，与把视觉投射到视皮质的视网膜细胞的相互位置关系相对应。从视网膜神经细胞向大脑皮质细胞发出的轴突束，保存了视网膜细胞的拓扑形态组织结构。这样，主要的视皮质表面就构成了一个视网膜表面的拓扑形态映射图。

称之为“拓扑形态映射图”，而不是简单地称为“映射图”，是因为视网膜细胞之间的**距离**关系一般**未经**保存。一般地，这样的图在度量上是可变的，它们就像是由橡皮制成，再以某种方式拉长。

有许多这样的映射图已获确认。所谓“视皮质”(17，18区)早已经有记载。躯体感觉皮质(3 区)的表层是身体的触觉表面的拓扑形态映射图。运动皮质(4 区)的底层是身体的肌肉系统的拓扑形态映射图。听觉皮质(41，42 区)含有频率空间的拓扑形态映射图。还有很多其他脑皮质区域，它们确切映射了什么，我们认识得还不够，但它们对远距离结构的拓扑形态表述是明白无误的。

这种一般的神经组织模式并不限于大脑半球的表面。更靠近脑中心部位的各种“灰质”神经核——例如上丘、海马和外侧膝状体——也表现出与此相同的多层拓扑形态组织的纵向联系结构。并非**每个组织**都是如此(例如，小脑就有很大不同，详细讨论见后)，但是上述模式是可在脑中发现的主要组织模式之一。

这样的模式为什么如此？它在功能上或认知上有什么重要性？这些结构是做什么的，又是怎样做的？通过对付第二个奥秘——感觉运动协调问题，我们就能够获得这些问题的可能解答。

3. 感觉运动协调

我以这个建议作为开始:纵向联系的分层结构,是进化过程为至关重要的一类问题提供的最简洁的解答之一,任何感觉运动系统,只要它超出最初的不成熟状态,都必须以某种方式解决这类问题。为了更好地理解这类问题,我们来看一个以略图表示的生物体,这是一个精心构思的简单物。

图 14-2b 是一个图示的螃蟹状生物体(图 14-2a)的平面图,这生物体带有两个可旋转的眼睛和一个可伸展的爪臂。如果要使这个装置对螃蟹有用,那么这个螃蟹就必须体现出它的眼角对之间在可食目标表现成三角关系时的某种函数关系,并体现出继之产生的肩部及肘部的角度,这样,它的爪臂才能据有一个与可食目标接触的位置。简单说,它必须能抓住它看到的东西,无论所见之物在什么位置上。

我们可以对所需的臂/眼关系的特点作出如下说明。首先,我们用二维感觉系统坐标空间或**状态空间**(图 3a)中的一个点来表示输入(眼角对)。输出(臂角对)也可用另一个二维**运动**状态空间中的恰当的点来表示(图 14-3b)。

我们现在需要一个函数,使我们从感觉状态空间中的任何一点到达运动状态空间中一个适当的点,这个函数将用上述方式使爪臂位置与眼睛位置协调一致。(下面,我将概述对相关函数的推导,这样就揭开了它的来历的奥秘,但是读者也可以不失理解地跳过代数运算。只要记住,我们正在推导一个可使我们从眼睛形态到达爪臂形态的适当的函数。)

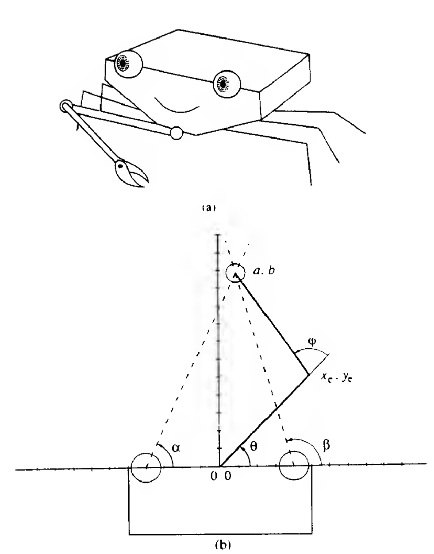
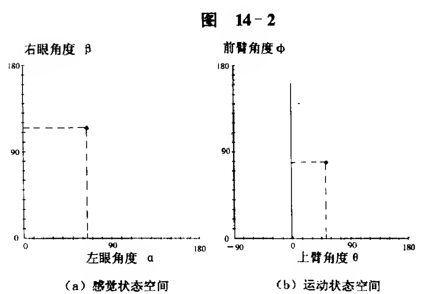


图 14-2



（a）感觉状态空间　　（b）运动状态空间

图 14-3

两个眼睛角度{α,β}确定出两条相交于所见目标的直线。该交点(在实数空间中)的坐标(a,b)由下式给出：

$$a = -4(\tan\alpha + \tan\beta)/(\tan\alpha - \tan\beta)$$

$$b = -8(\tan\alpha \cdot \tan\beta)/(\tan\alpha - \tan\beta)。$$

爪臂顶端必须与这点接触。假定前臂和上臂都具有7个单位的固定长度,因而肘部只能位于半径为7个单位的两个圆的交点上:一个圆心在(a,b)处,另一个圆心在(0,0)处,也就是上臂从螃蟹身体上伸出的那个地方。求出这一相应的交点之后,实数空间的肘部坐标(x_e,y_e)可给出如下：

$$x_e = ((49-((a^2+b^2)^2/4b^2)\cdot(1-((a^2/b^2)/((a^2/b^2)+1))))^{1/2}+(((a/b)\cdot((a^2+b^2)/2b))/((a^2/b^2)+1)^{1/2}))/((a^2/b^2)+1)^{1/2}$$

$$y_e = (49-x_e^{\ 2})^{1/2}。$$

实数空间中的三点(a,b),(x_e,y_e),(0,0)确定出爪臂的位置,其上臂和前臂的角度{θ,φ}最后给出如下:.

$$\theta = \tan^{-1}(y_e/x_e)$$

$$\varphi = 180-(\theta-\tan^{-1}((b-y_e)/(a-x_e)))。$$

这些就是爪臂在运动状态空间中所需要的坐标。读者会注意到,合在一起的产生这些坐标的函数是相当复杂的。

无论复杂与否,如果在计算机屏幕上画出这个螃蟹,使它爪臂的最终位置(由计算机画出作为输出)就是它的眼睛位置(由我们输进作为输入)的特定的函数,那么这就构成了一个非常有效的和举止得当的感觉运动系统,特别是在我们编写如下的控制程序时。

设该程序使得螃蟹爪臂弯曲地靠在它的胸前(在$\theta=0°,\varphi=180°$处),直到某个适当的刺激对准两眼的中央凹为止。然

后，让它的爪臂从初始状态空间位置(0°,180°)，沿着运动状态空间中的一条直线，向运动状态空间中计算好的目标位置运动。这就是在实数空间中爪臂的顶端与眼睛的三角测量点相接触的状态空间位置。这种安排产生出一个适度的仿真系统，无论它看到什么东西，只要在它爪臂的可达范围之中，就可以准确无误地抵达(图 14 - 4a - c)。

前面所列六个方程中关于螃蟹感觉运动变换的代数表述，并没有提供螃蟹的总性能的直觉概念。几何学表示的启发意义要大得多。下面我们来考察从螃蟹感觉状态空间(图 14 - 5a)的活动部位，到它的运动状态空间(图 14 - 5b)的正交网格的投影，它是由前面提出的函数形成的。也就是说，对于在感觉网格中显示的每一个点，我们都已经测绘出它在运动网格内的相应的爪臂位置。

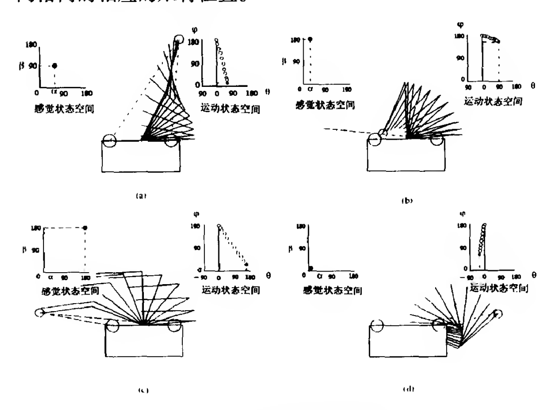
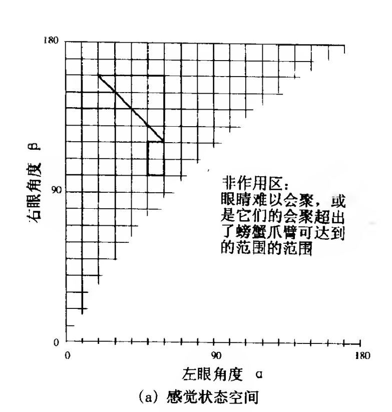
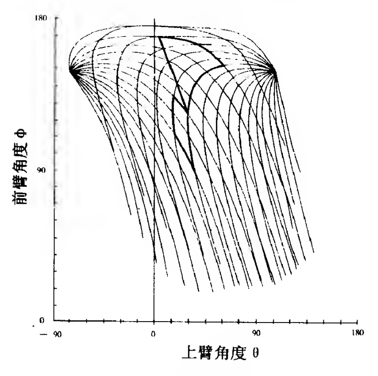


图 14 - 4 行动中的螃蟹爪臂



(a) 感觉状态空间

感觉网格在运动状态空间的投影



(b) 运动状态空间

图 14-5

这里，我们一眼就能看清向运动空间作投影后的感觉空间的垂直线和水平线的变形。感觉空间的拓扑特征得到保存，但是它的度量特性没有得到保存。我们看到的是系统的**坐标变换**。（用粗线条画出三角形和矩形，只是为了帮助读者确定在变形网格和未变形网格中的对应位置。还应该注意，图 14－5a 的左边界或 β 轴收缩成图 14－5b 中的左辐射点，而图 14－5a 的顶边界收缩成图 14－5b 中的右辐射点。）

4. 坐标变换：它的物理实现

上述变换为有效的和现实的感觉运动行为提供了根据。但是一个真实的神经系统怎样才能计算这样一种复杂的坐标变换呢？期望它像我们的计算机模拟那样一步一步地去计算这种复杂的三角函数是不现实的。然而，如果已知它们有精致的感觉运动协调，那么生命系统**必然**以某种方式计算像这样的变换和另外一些更复杂的变换。它们怎样能做到这一点呢？

图 14－5 给出了一个格外简单的方法。如果我们假定螃蟹含有对它的感觉状态空间的内在表述，以及对它的运动状态空间的内在描述，那么下述安排会实现所希望的变换。设螃蟹的感觉状态空间用一个由携带信号的纤维构成的物理网格来表示，这网格是像图 14－5b 所示那样的在实数空间中的度量变形网格。再用第二个纤维网格以未变形的正交排列方式表示它的运动状态空间。把第一网格置于第二网格之上，并且用大量从感觉网格中的坐标交点向下延伸到底层运动网

格中最近的坐标交点的纵向短纤维将它们连接，如图 14－6 所示。

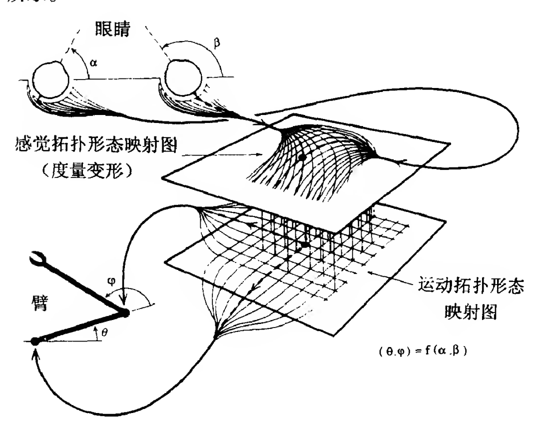


图 14－6 用相邻拓扑形态映射图所作的坐标变换

假定感觉网格的纤维接收到来自眼本体感受系统的输入，使每个眼睛的位置刺激上层（变形的）网格中唯一的纤维。左眼激活来自右辐射点的一根纤维，右眼激活来自左辐射点的一根纤维。于是，通过同时发生在上层网格中恰当坐标交点上的刺激，联合的眼睛位置就得到表示。

在上层映射图中该点的下面，存在着运动网格中的唯一交点。我们设定，在联合激活时，正交运动纤维中的这一交点对将导致爪臂据有适应于发出这一运动信号的特定运动坐标交点的位置。

幸运的是，这些映射图中的相对度量变形已经以对应方

式在上下映射图中设置了恰当的点。现在我们只需假定,感觉网格和运动网格之间的纵向联系起着像“与门”或“阈开关”那样的作用,这样,当相关感觉交点同时被它的两根相交感觉纤维刺激的时候,一个信号就会被准确地顺着纵向联系向下送往运动网格。这样的系统将计算出所希望的坐标变换,其精度只受到这两个网格的粗细程度以及它们的纵向联系的稠密程度的限制。我把这样的系统称为**状态空间分层结构**。

值得注意的是与这一安排方式的功能特性直接有关的三个方面。首先,尽管有局部破坏,它仍然能保留部分功能。任一网格中的一个小损伤只产生局部的运动障碍(从损伤处顺流而下的纤维不活动性造成的两个永久性“盲区”),这通常可由身体位置的移动来补偿(使目标的状态空间位置离开这个盲区)。

其实我们甚至还可以做得更好一些。用一个小棒替代图示螃蟹眼睛(图 14 - 6)背后的凹座,这样,刺激的就不是一个细胞,而是一组相邻的本体感受细胞。然后,通过激活中心位于“正确”纤维上的一束纤维,就可以将每个眼睛的位置表示在上层映射图中。于是,通过一个分布着刺激的**区域**,即一个中心位于“正确”地点的区域,**联合的**眼位置就被记入上层之中。这将在底部网格中引起一个对应的刺激区域,并因而将一束刺激送到每一块肌肉。如果把这些肌肉连接起来,使它们据有一个对那个分布式信号内的“中间”纤维来说是恰当的位置,即使这一特殊纤维碰巧是不活动的也没有关系,于是恰当的运动响应将会随之而出现,即使分层结构已经分散地损失了数量相当多的细胞,也不受影响。像这样的系统,即使细

胞广泛损坏，其功能也可得到保持。由于细胞受损，感觉运动协调的质量会有某种程度的降低，但是一个大致恰当的运动响应仍然会出现。

其次，这种系统非常非常之快，虽然它是以具有生物传导速率（$10 < v < 100$ 米/秒）的纤维传导的。在像螃蟹那样大小的动物中，传导通路全长小于 10 厘米，这种系统产生运动响应的时间远远低于 10 毫秒。在上述螃蟹模拟中，我的微机（在软件里做三角运算）为了得出屏幕上的运动响应，花费的时间是这一时间的 20 倍，而微机的传导速率与光速是同一数量级的。显然，状态空间分层结构的大规模并行构造使它在速度上赢得很大的优势，虽然它是由慢得多的部件组成的。

第三，在整个运动活动范围内，螃蟹的协调性能是不均匀的，因为在感觉网格变形最大的区域中，感觉记录中的小错误也会产生出运动响应（再参阅图 14－5b）中的大错误。因此在眼睛之间的封闭区域中，以及最左边和最右边的区域中，螃蟹的协调性最差。

这三个功能特性都符合生物学上的现实特征。而分层结构的生物学现实特征还进一步表现在这个方面：想象这种系统的形成，是相对容易的。不同的层次可以对不同的化学梯度作出限制，因而可以引导不同的形态发生过程。因此，不同的拓扑形态映射图可以出现在紧邻的层次中。但是如果已知映射图是如此紧邻的，并假定它们有恰当的变形，那么将这些层次连接起来产生出一个函数系统的问题，就变得很平常了：其解答只不过是形成与这些层次大体上正交的传导元件。

不同的动物有着不同的目标定位手段，和实现与目标接触的不同的运动系统，但是它们全部会面对同一问题，即感觉

状态空间中的位置与运动状态空间中的位置协调一致的问题，而本文概述的解答类型，其性质显然是相当普遍的。事实上，据推测，不同的动物子系统通过坐标变换实现协调的范围，已远远超出了基本感觉运动协调这种明显的情况，同时正像我们后面将要看到的，在执行较高级的认知活动时，同样的策略可能仍是有用的，甚至是必不可少的。这里要强调的是，对于实现**任何**二维到二维的协调变换来说，无论它在数学上有多么复杂，无论坐标轴对脑可以代表的是何种特征——内部的或外部的，抽象的或具体的，状态空间的分层结构总是构成一个简单的、生物学上现实的手段。只要这个变换能够用图形表示，分层结构就能计算它。前面解决的感觉运动问题，只是正在使用的一般性技术中的一个简洁明了的例子。

图 14－6 中的互连映射图系统，除了它的函数实在性而外，还与已知的典型分层大脑皮质物理结构，包括分布在整个大脑表面的许多拓扑形态映射图，具有易使人产生联想的相似性。在所有这些区域中，输入所在的位置是某个已知的细胞层，该细胞层多次体现出这样或那样的度量变形拓扑形态映射图。而输出所离开的区域则存在于不同的层内，第一层与该层之间有大量纵向联系。

因此，我提出这种假设：大脑皮层内那些分散的映射图，以及许多亚大脑分层结构都从事于从一个神经状态空间中的一些点到另一个神经状态空间中的一些点的坐标变换，其做法是使纵向联系的度量变形拓扑形态映射图直接相互作用。它们的表述方式是状态空间位置；它们的计算方式是坐标变换；而这两种功能在状态空间分层结构中同时得到实现。

【（1989 年增补）这个假设的第二部分，特别是第三部分，

现在看来几乎肯定是错误的，至少在对大脑皮质的解释上是这样。特定大脑皮质区域的特定皮层中的细胞群体确实是在对状态空间的位置进行编码，但采用的是全体细胞均处于激活水平的全局模式，而不是对最强的细胞激活所作的狭隘空间定位。另外，从这一层到邻近细胞层的轴向投射，的确实现了从一个状态空间到另一个状态空间的变换，但是其方法是下文§6中概述的"矩阵乘法"式的变换，而不是转移相互变形映射图之间的一个激活热点。这种拓扑形态映射是许多皮质区的特点，而现在则表现为带有深层编码策略的少数人工制品：参阅§6中对多维矢量编码的解释。然而，作为对上丘的一种可能的解释，以及作为对有关计算思想的介绍，本节的讨论仍然是富有启发性的。】

我无法举出一个已知的单个大脑皮质区，对它来说这种功能假设是正确的。大脑皮质映射图的解码是一件刚刚开始的事情。明显成功的例子，屈指数来最多不过十几个，并且一般限于表面的大脑皮层。然而，有一个重要的皮质下区域，它的上层映射图和下层映射图已至少被部分解码，并且它的确表现出图14-6描绘的那种一般模式。

上丘在种系发生上是一个非常古老的分层结构（图14-7a），位于上中脑背侧。除别的功能而外，它还使我们熟知的条件反射得以维持，眼睛则靠这种反射做无意识扫视，以便在中央凹形成（=直视）任何进入视网膜的、离开中央凹中心的突然变化或运动。在黑暗的电影院中，当前排左边某个人突然用火柴或打火机点燃香烟时，我们都曾有过这种体验。剧院中的每只眼睛在返回银幕之前，都用一个飞快的扫视去注视这一短暂的刺激。这就是上丘在起作用。这种现象有时

被称为“视觉获取反射”,是完全恰如其分的。

在人类和较高级的哺乳动物中,上丘位于大脑半球后方,是仅次于更重要的纹状皮质(布罗德曼图中 17 和 18 区)的视觉中枢,但在像蛙或蛇这些没有任何重要脑皮质的较低级动物里,上丘(或**视觉顶盖**,这是在低级动物中的叫法)则是它们主要的视觉中枢。然而,即使对哺乳动物来说,它也是一个重要的视觉中枢,它的作用大致如下:

上丘(以下简称 SC)的最外层接收直接来自视网膜的投射,同时它构成视网膜表面的一个度量变形的拓扑形态映射图(图 14－7b)(Schiller 1984; Goldberg and Robinson 1978; Cynader and Berman 1972; Gordon 1973)。纵向元素把该层连接到 SC 的最深层上。这些纵向联系看上去是由一根链组成的,链上有两三个短轴突中间神经元,穿过两个介人层逐步走向深层(Schiller 1984: 460, 466),详述见后。同样,某些深层神经元的树突看上去是直接向上伸入视觉层的,形成与视细胞的突触连接(Mooney et al. 1984: 185)。最深层的神经元发出的传出轴突,经由两条不同的神经通路,分别与控制眼睛垂直运动和水平运动的眼外肌对相连(Huerta and Harting 1984: 287)。

令人感兴趣的是,这一基础运动皮层也体现出一个拓扑形态映射图,它是一个表示眼睛肌肉可收缩位置变化的状态空间映射图(Robinson 1972: 1800)。用一个电极在最深层的某点形成一个微刺激,就会使眼睛执行一个具有特定大小和方向的扫视,这一扫视将中央凹锁定在视网膜细胞以前所占据的位置,视网膜细胞则向上丘最顶层中即刻被覆盖的视觉细胞层投射(Robinson 1972; Schiller and Stryker 1972)。换句话

说，两个映射图中的相对度量变形已经使上图和下图中的适当的点处于相对应的位置。（这意味着，在图 14－7b 中看到的“变形”，就其本身来说，不应看作是这种状态空间分层结构假设的根据。起决定作用的是映射图的相对变形。）

最后，在最上层视觉映射图中，视网膜产生的任何足够强度的刺激，通过适当的纵向元素向下传送到运动映射图上，在这里产生一个快速扫描，该扫描的大小和方向恰好对准中央凹并与刺激它的外界刺激相对应。这样，SC 在结构和功能两个方面都可以看作是图 14－6 所示模式的实例。

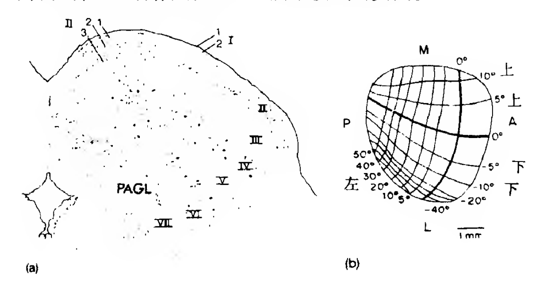


图 14－7

（a）猫的上丘尼氏染色法剖面投影图，说明分层组织。图中的点表示上丘神经元。（Kanaseki and Sprague 1974 允许重印）

（b）视网膜行为模式映射图：正交坐标中猫的右上丘表层上视觉半扫描场的度量变形拓扑形态映射图。M＝中间；L＝侧面；A＝前面；P＝后面。（据 Schiller 1984 改编）

人们也许预料到，动物的分层结构是用上层映射图中的刺激区域，而不是单个的点，来为视网膜刺激位置编码的，因而如前面解释的那样，在遇到小的损伤和分散的细胞死亡时，其功能仍能保持不变。SC 中的活动的确表现出这种模式

(McIlwain 1975; 1984; 268)。图 14－6 表示的模型也预示出，由上丘分层结构内部的各个点上的微刺激所引起的运动响应的大小和方向，只是每一映射图中刺激发生地点的函数，而不是刺激大小的函数，也不是两个映射图之间纵向位置的函数。实验已经得出这种结果(Robinson 1972; Schiller and Stryker 1972)。正如所表现的那样，SC 是一个真正的感觉运动坐标变换器，与所讨论的那种大致相同。它使中央凹对准正在变化或移动的视觉目标，所用方法与图示的螃蟹脑皮质指向处在三角形位置上的目标的方法基本相同。

这里有必要提醒一句，刚才所作的解说并不是对 SC 的全部复杂性而言的。在哺乳动物中，特别在较高级的哺乳动物中，SC 是一个较大调节系统中的一个严密整合的部分，该系统包括来自视皮质和前方视野的输入，以及对颈部肌肉的输出。整个系统的功能特性比前面简要说明的情况更加富于变化，也更加精细，对其分类的工作还正在进行中(Mays and Sparks 1980; Schiller and Sandell 1983)。以上讨论至多可以看作是对 SC 主要功能或较基本功能的解释。

对这些例子——螃蟹的"脑皮质"，以及上丘——有所了解之后，合理的做法是把重点放到其他许多按拓扑形态方式组织起来的、散布在整个脑上的多层脑皮质区上，同时提出它们有可能实现何种坐标变换的问题。这里很重要的一点是，应当意识到，我们寻求进行解码的拓扑形态映射图不需要是，一般也不是某种有明显解剖学特点的东西例如视网膜表面或皮肤表面的映射图。在更多的情况下，它们是某种**抽象状态空间**的映射图，其维度的意义可能是从因果性出发的观察者所难以理解的，虽然它们对脑来说在功

能上是极为重要的。这种抽象映射图有两个很好的例子——蝙蝠听皮层中的**回声延迟**图和猫头鹰下丘中的**双耳差异**图(Konishi 1986)。

在图示的螃蟹脑皮质的例子中,一个外部系统(眼)的角度状态,直接映射到另一个外部系统(带关节的爪臂)的角度状态上。而在具有任何复杂性的动物中,我们都可以预期有一个由相互作用的一些内部系统组成的长链或层级体系,这些系统是另一些内部系统的输出的映象,而它们的输出则驱动另一些内部系统的活动。我们要理解这种映象,必须理解总系统中其他映象的功能。

所有这一切表明,脑可能具有的拓扑形态映射图,比至今已确认的,甚至是推测的,还要多得多。脑无疑具有极其丰富的以拓扑形态方式组织的区域,最近的工作又将已知的感觉相关映射图的数量大为扩充(Merzenich and Kaas 1980; Allman et al. 1982)。所有这一切都进一步表明,在试图理解许多以拓扑形态方式组织的脑皮质区的重要性时,如果把它们作为抽象的可是在功能上相关的状态空间的映象来处理,我们将会取得更大的进步。

在神经科学家中间一直有一种倾向,即把术语"拓扑形态映射图"局限于一些反映物理世界或感觉系统的某一简单方面的神经区域,比如视网膜或皮肤表面。这是令人遗憾的,因为脑在它构成什么东西的映象方面没有任何理由表现出这种偏好。抽象状态空间正如具体的物理状态空间一样,是可以映射的,而脑对于什么是重要的,事先确实并不知道。我们倒是应当期望脑对于功能上重要的东西逐渐形成一些映象,而这东西往往是抽象状态空间。

5. 多于两层的脑皮质

在讨论上述分层机制的生物学实体时，我们来看看这种不同意见：我们模型中的脑皮质只有两层，而典型的人类大脑皮质为六层，如果算上细分的亚层，某些区域也许有八或九层。它们的作用是什么呢？

这样的增加层的功能是不难了解的。让我们再回到上丘，它对这里的众多可能性作出了一种解释。在某些动物的SC的视觉图和运动图之间，存在着一个或两个中间层（再参阅图14-7）。它们看来构成了一个听觉图和（或）一个躯体感觉图（面部图或触须图），其功能又是调整眼睛中央凹的位置，这次是转向突发的听觉和（或）躯体感觉刺激源（Goldberg and Robinson 1978）。正如预料的那样，这些介入的映射图，每一个都发生了度量变形，变形的结果是与运动图之间形成粗略的坐标"记录"，因而相互之间也是如此。总之，这种精致的三或四层拓扑状态分层结构，构成了一种多模态的感觉运动坐标变换器。

多层结构还有进一步的优点。很显然，数种不同模态的映象，由于它们在一个"棒状分层结构"（club sandwich）内的总的记录中的恰当变形和定位，为交叉模态的整合和比较提供了最有效的方法。例如，在SC中，这种多模态的安排适合于对联合接收到微弱的、但在时空上一致的听觉和视觉刺激产生运动响应，这些刺激孤立出现时总是低于运动响应的阈值的。例如，来自某一方位点的微弱声音，也许太弱，不能促使

眼睛进入中央凹扫视状态，而来自某一方位点的一个细小运动也许同样是不起作用的；但如果声音和运动这两者来自同一方位点（因而在SC中沿同一个纵向轴编码），那么它们在同一时刻的结合就确实足以使运动皮层对眼睛作出恰当的指示。梅雷迪斯和斯泰因（Meredith and Stein 1985）最近的研究成果是对这一预测的有力支持。

进一步的探索揭示，多层分层结构可以明显地促进复杂的认知功能。在先前关于这些问题的一篇文章中（Churchland 1986），我已经说明了一个三层状态空间分层结构如何能以适当的方式为运动目标的轨迹编码，并进行投影，以确定螃蟹捕捉正在飞动的运动目标所需要的爪臂位置。很显然，多层的脑皮质会具有相当多的优越性。

6. 超出状态空间分层结构的情况

在上面研究的例子中，所具有的输入状态空间和输出状态空间都毫无例外地是二维的。其原因在于这一事实：所需坐标变换可以通过一对相邻的片状的映射图来实现。但是，那些所含子系统分别具有多于两个的参数的情况如何呢？从n维输入空间到m维输出空间（这里n和m是不同的，并且两者都大于2）进行坐标变换的情况又如何呢？让我们考虑一下，例如，带有三个以上关节的肢体的合成角度的协调问题，以及数个这种肢体互相协调的问题。或者考虑一下共同控制这种肢体的更大数量的肌肉的协调问题。一旦考察了那些真实动物平常面对的、并已经解决了的问题，人们就会感到

其中许多问题远比简单的二维到二维变换所能表示的要复杂得多。

在这些较复杂的问题中,有一些也许可以通过把它们划分成一组较小的问题来解决,这些较小问题最终能由一组不同的二维状态空间分层结构来处理,其中每一个分层结构为较大问题的某部分或某方面编址(关于这一思路的某些专门建议见 Ballard 1986)。脑的分层脑皮质的优越性无疑会促使我们沿这一路线作思考。但是这种解答,即使是近似的,一般说来也不能得到保证。如果脑是以平常方式来处理这些更高维问题的,它就特别需要某些高于状态空间分层结构的机制。

派利欧尼斯和利纳斯曾概述了一种适合于这种任务的机制,并在小脑内部找到了实现它的证据,给人留下深刻印象。小脑是脑后部的大结构,就在大脑半球的底下。它的主要功能最初是从损伤研究中发现的,是协调复杂的身体运动,如准备晚餐或打篮球时所表现的那样。它表现出的神经组织与大脑半球的组织有很大差别,这种组织的重要性可能通过派利欧尼斯-利纳斯的说明而变得清楚起来。

为了说明这种较一般的坐标变换机制,让我们考察一个四维输入系统,其输入 a,b,c,d 被变换成三维输出系统的值 x,y,z。像以前那样,输入和输出可分别被看作适当的状态空间中的点。因为它们是 n 维的,所以每个量都可被看作一个矢量(其起点位于相关状态空间的原点,而其端点位于以 n 维表示的点上)。

用于从矢量到矢量的系统变换的标准数学运算是矩阵乘法。这里,正是**矩阵**体现出或实现了所要的坐标变换。为了弄清这是怎么完成的,我们来看看图 14-8 的矩阵,该矩阵有

4行3列。要将这个矩阵与输入矢量{a,b,c,d}相乘,就把p_1乘以a,p_2乘以b,p_3乘以c,p_4乘以d,然后把4个结果加起来得出x。然后对第二列重复这个过程得出y,再对第三列重复这个过程得出z。从而得出输出矢量{x,y,z}。

$$\{a,b,c,d\}\cdot\begin{bmatrix} p_1 & q_1 & r_1 \\ p_2 & q_2 & r_2 \\ p_3 & q_3 & r_3 \\ p_4 & q_4 & r_4 \end{bmatrix}=\{x,y,z\}$$

图 14-8

这种代数运算可以通过图14-9的神经分布,用物理方式相当简单地予以实现。右边的平行传入纤维分别把一串电化学“脉冲”送往那些等候着的树突树。数字a,b,c,d表示四根纤维各自的瞬间脉冲频率高于(正数)或低于(负数)某一基准脉冲频率的量。例如,最上面的传入纤维分别与三个传出细胞形成突触,各产生一个刺激联系,使细胞胞体去极化,并把一个脉冲沿着它的纵向传出轴突向下送出。每一个细胞的脉冲发射的输出频率取决于:(1)它从所有传入的突触联系接收到的输入刺激的单个频率。(2)每个突触联系的**加权值**或**强度**,由突触的布局和它们的截面积确定。这些强度值分别由图14-8中矩阵的系数来表示。**于是,神经的相互联结特性实现了这个矩阵。**图14-9中的三个细胞分别把它们接收到的刺激“加”起来,并沿着传出轴突向下发出一串适当的脉冲。这三个输出频率不同于三个传出细胞的背景频率或基准频率,相差量或正或负,这些量对应于输出矢量{x,y,z}。

注意,对于状态空间分层结构来说,信息编码是一个神经事件的空间定位的问题。与此不同,对所讨论的矩阵乘法的

计算类型来说，输入和输出变量是由相关通路中的几组脉冲频率来编码的。前一系统使用“空间编码”，后一系统使用“频率编码”。但两种系统所从事的都是状态空间位置的坐标变换。

图 14-9 的例子是一个 4 乘 3 矩阵。但是显然，无论数学运算，还是它的物理实现，都不受任何维数限制。在原理上，派利欧尼斯-利纳斯的联络矩阵可以实现维数达到数千个甚至更多的状态空间的变换。

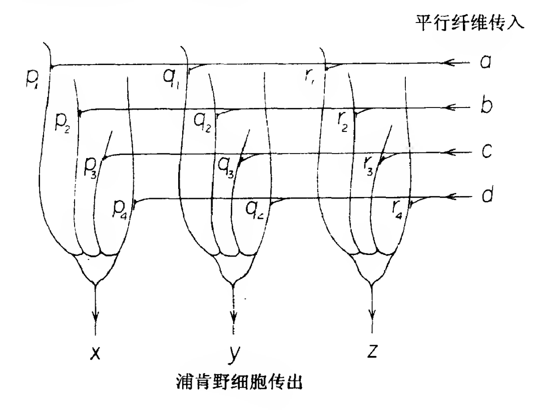


图 14-9

图 14-9 示意的结构非常接近于小脑中发现的微细结构格局(图 14-10)。(参阅 Llinas 1975，可知有关小脑构造体系的概况。)水平纤维在那里称为**平行纤维**，它们是从较高级的运动中枢传入的。密集的垂直位的细胞则称为**浦肯野细胞**，它们的传出穿过小脑核到达运动终末。事实上，由于对小脑

精密规范的构造体系的观察，同时也由于试图通过在计算机中建立小脑的大规模物理联系，以重新建立小脑的功能特性，派利欧尼斯和利纳斯一开始就倾向于这种观点：小脑所做的工作，是将一个神经多维空间中的矢量系统地变换为另一个神经多维空间中的矢量(Pellionisz and Llinas 1979)。

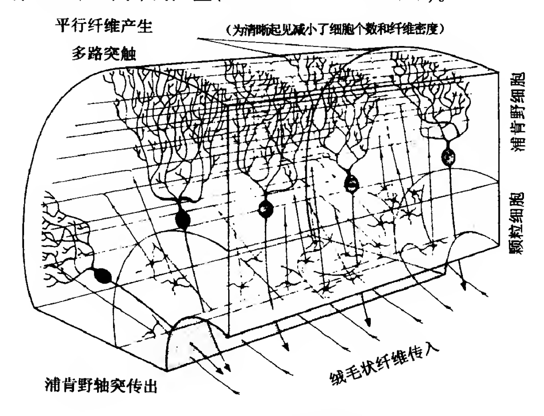


图 14-10　小脑剖面示意图

了解了该问题的这个观点之后，特别是由于我们不能指望脑本身仅局限于笛卡尔坐标，张量计算自然就可以作为用来对这类事物编址的框架。在至此讨论过的例子中，沿着相关状态空间的任何一根轴的位置变动，与沿任何别的轴的变动无关，但是这种独立性并不表示非正交轴状态空间的特征。的确，在派利欧尼斯和利纳斯看来，该方法包括了非笛卡尔多维空间在内的这种普遍性，是他们这一解释的最重要的特征之一，除最简单的协调问题而外，这一特征对理解所

有问题来说都是必不可少的。这里我不能继续讨论这一特征了。

对图 14－9 中神经矩阵的最后三点总结。第一，它不必局限于线性变换的计算。单个的突触联系能表示广泛范围内的任何功能特性。它们无需只是简单的乘法器。所以它们能以协同方式计算多种多样的非线性变换。第二，神经矩阵具有与状态空间分层结构同样的非常之快的速度。第三，如果给出的是大型矩阵和(或)细胞的庞大组织，即使它们分散地损失一些细胞成分，这样的结构也会表现出对功能的保持。

这些简要的评述当然不是对派利欧尼斯和利纳斯极其丰富的研究成果的全面评价，同时我也没有提出任何批评意见。(后者见 Arbib and Amari 1985。对他们的答复见 Pellionisz and Llinas 1985。)至于更深入的讲解，读者必须查阅文献。这一节的主要内容是说明，本文前面提到的一般性功能的图式，即用状态空间位置表述和由坐标变换计算的图式，即使在表述作业和计算作业超出二维情况时，也不会遇到实施的困难。相反，脑包容的神经机构可以完美地适合于维度很高的情况。这样，我们至少具有两种已知的执行坐标变换的脑机制：专门用于二维情况的状态空间分层结构，以及不论对何种维度都适合的神经矩阵。

7. 状态空间的表述力量

到目前为止，讨论一直集中在状态空间坐标变换的令人印象深刻的计算力量上，以及这种活动的可能的神经实现

上。但重要的是，要充分意识到神经状态空间的同样强有力的**表述**能力。有 n 个不同变量的复杂系统的整体状态，能够由抽象的 n 维状态空间中单个的点来表述，这是很经济的。而这样的状态空间点在最简单的情况下，能够由只有 n 根不同纤维的系统中的 n 个脉冲频率的特定分布以神经方式实现。此外，状态空间表述还体现出它之中不同的可能位置之间的**度量**关系，从而体现了对这样表述的不同项目之间的**相似性**关系的表述。我们用五个例子来说明这些论断，它们都是真实的，而其中三个例子提出了一些哲学家们所熟悉的问题。

对心理状态的任何神经生物学简化来说，大家公认我们知觉的性质特点提出了一个特别棘手的问题（见 Nagel 1974；Jackson 1982；Robinson 1982）。在呈现于意识中的主观上可辨别而"客观上不可描绘的"感觉特性中，的确很难看到为简化目标留有多少位置。

即使如此，为找出这一领域的有序性而不是神秘性所作的坚韧不拔的努力，还是揭示出了大量的可以表达的信息。例如，我们都赞成，我们视觉感觉的"颜色"感觉特性自行排列在一个连续体中。在这些特性所处的连续体中，存在相似性关系（橙色与红色相似），相对相似性关系（橙色与红色比与紫色更相似）和中间性关系（橙色处于红色和黄色之间）。同时，还存在着不定数目的穿过连续相似颜色的不同的"通路"，可使我们从任何已知颜色达到一个不同的颜色。

对此，我们可以补充说，那些患有这样或那样不同类型的色盲的人，其实体现了颜色感觉特性的一个大大**简化**的连续体，至少是以部分可说明的方式简化的连续体（它不能表示红

绿对比,或蓝黄对比,等等)。一个问题是在一个给定的感觉模态中出现的感觉特性的相对种类数,这个问题提出了这种看法:在我们熟悉的五种感觉模态间,存在着值得注意的变化。例如,虽然可辨别的颜色感觉的种类非常之多,但是可辨别的味道感觉的种类甚至更多,而可辨别的气味感觉的种类还要多。这种感觉模态间的变化,使我们进一步想到物种间的假定变化,以犬类的超常辨别能力为例,它们单凭嗅觉就能辨别出这个星球上 35 亿人口中的任何一个人。就含有更多种可辨别的感觉类型而言,人们足以推测,由于某种原因,犬的嗅觉感觉的连续体比人类的"大"得多。

于是,我们得到一些关于主观感觉特性多样性的普通事实,这些事实可尝试由对心灵的简化说明来作出解释。要做到这一点,就必须用神经生物学术语,以某种具有启示作用的系统方式,对这些事实进行重构。(有关通过理论的同一性的本质和理论相互之间简化的本质的一般性说明,见 Churchland 1985,1979。)现在来探讨这种做法的可能性。对几种相关的感觉模态来说,生理心理学家和认知心理学家已草拟了这样的说明的大纲,而状态空间表述在所有这些大纲中起着突出的作用。

首先考虑 E·兰德提出的抽象三维"颜色立方体"(图 14-11),在人类可辨别的数百种颜色中,每一种颜色都在这立方体里占有唯一的位置或一个小体积(Land 1977)。每一根轴表示眼/脑以我们的视锥细胞有选择地响应的三种波长之一构造所见对象的客观反射的情况。两种颜色只有当它们在这一立方体内部的状态空间位置相互接近时才是十分相似的。只有当两种颜色的状态空间位置相距很远时,它们才是

不相似的。我们甚至能够说出相似的程度，以及计算相似性时所沿的维度（也见 Zeki 1983）。

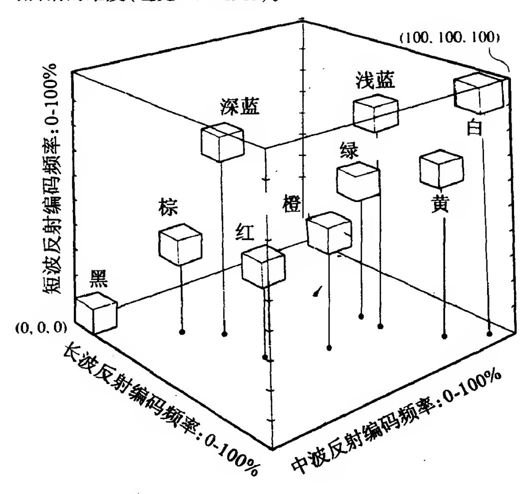


图 14－11 颜色状态空间

如果人脑的确拥有这种状态空间的内部实现方式，那么它就以很低的代价换取了很大的表述力量。例如，如果沿着兰德的颜色状态空间的每一根轴，我们本来的辨别能力只有 10 个不同的位置，那么一个三维系统就能够足足表示 10^3 个不同颜色。即使如此，这也低估了我们的能力，所以假定沿轴线有 10 个单位的辨别能力可能太低了。不管怎样，用它来解释我们广泛的辨别力量是没有问题的：在兰德状态空间里的人的辨别力，作为沿每个轴的辨别力的三次幂而迅速增大。

我们的周围神经系统机构确实证实了这个一般性假设。所有颜色知觉都是由恰好三种视网膜视锥细胞的输入产生的。

所有这些提出了这一假设：任何特定颜色的视觉感觉实际上等同于某一具有三元一组形式的脑系统中的特定的三重脉冲频率。如果情况果真如此，那么两种颜色感觉的相似性就仅仅表现为它们各自状态空间位置的近似度。性质上的"中间性"成了状态空间的中间性。当然存在着不定数目的连接任何两个状态空间点的连续的状态空间通路。显然，我们可以再构想一个如图 14－11 那样的立方体作为内部"感觉特性立方体"。我们只要认为每个轴表示着反射信息所通过的三个内部通路之一的瞬时活动水平或脉冲频率即可。

最后，如果先天不幸使一个人失去了正常的三个通路之一，那么他或她的感觉特性空间就会缩降为三种可能的二维空间之一，至于是哪一种，则取决于三个轴中的哪一个变成无效的了。一个可预见的特定的颜色辨别力缺陷，应与每一缺失相伴而生。实际情况也是如此。有三种基本的色盲类型，每一种对应于视网膜视锥细胞三种类型之一的损失。这里我们完成了对于以纯粹简化方式解释一个感觉特性领域的概述。

我们的味觉系统看来运用了类似的安排，不过这里的状态空间维数表现为 4，因为这是口中不同类型的味觉感受器的数目。所以人们推测，人的每一可能的味觉都是四维味觉状态空间里某处的一点。或更准确地说，它们是四个专用通路中的四重脉冲频率，这些通路把来自味觉频率分布感受器的信息送往脑的其余部位。如果我们沿着每个轴的辨别力与颜色空间里的辨别力相当（每个轴上有 10 个单位以上），这就

意味着不同味觉的种类数将比不同颜色感觉的种类数大约大一个数量级。看来也确实如此。味觉的这种状态空间研究法是以**交叉纤维模式理论**出现在神经科学文献中的(Bartoshuk 1978; Smith 1983; Pfaff 1985)。

对味觉空间的这一认识使得我们可以确定地说出"它像什么",是老鼠还是猫。与人类一样,老鼠和猫也是哺乳动物,并且它们也拥有四通道的味觉系统。然而,有一种差别值得一提。四个通路之一——常常记为"苦味"通路,因为对苦味的四值编码需要这一通路中的高级活动——显示出这三个物种间的不同灵敏度。同人类相比,老鼠的这一通路可引发的活动范围显得较窄(较低的辨别力);猫的这一通路的活动范围则显得较宽(较高的辨别力)。

现在来看看人类对糖(蔗糖)的味觉和对糖精的味觉的细微对比。一般是糖受欢迎,因为糖精有轻微的苦味。上述关于老鼠和猫的情况说明:老鼠的苦味差别将小于人的这一差别,而猫的差别则大于人的。即糖精的味道对猫会比对人更苦些。这也就是上述情况所说明的。实际情况是老鼠的选择行为不能辨别出糖和糖精:它们没有区别地两样都吃。而猫却不同,它吃糖,而拒绝吃糖精(Bartoshuk 1978)。

与此相同的一般性质的说明,也适合于我们的嗅觉系统,该系统有六个或更多的不同类型的感受器。六维空间具有更大的容量,因而可能有甚至更大的辨别本领。在轴向有 10 个单位辨别力的情况下,六维空间可能有 10^6 种气味的辨别力。我们只要设想一个七维嗅觉空间,轴向的辨别力三倍于人类——这差不多就是狗所肯定具有的辨别力,那么我们考虑的就是一个具有 30^7 个或 220 亿个辨别位置的状态空间!了解

到这一点，犬靠嗅觉去辨别这个星球上 35 亿人中的任何一个人的能力就不再显得有什么神秘了。

在讨论听觉感觉特性的复杂情况时，我既没有这种空间，也没有这种认识，但这里提出的状态空间方法仍被认为是富有启发性的(见 Risset and Wessel 1982)。由于研究人员的不同和所涉及的模态的不同，这状态空间方法有不同的名称："多元分析"、"多维扫描"、"交叉纤维模式编码"或"矢量编码"等等。但是，它们都是同一个东西的化身：状态空间表述。

显然，这种认识感觉特性的方法受到来自理论和实验两方面的推动，它也是对我以前在一篇关于感觉特性的本体论地位的文章(Churchland 1985)中提出的简化立场的支持。特别是，它提出了一种有效的表达工具，可以表达那些被断言为无法表达的东西。人们当前视觉中"难以用语言表达的"粉红色，也许可以充分而准确地表达为一个相关的三元一组的大脑皮质系统中的"95Hz/80Hz/80Hz 的频率调和"。由冒牌的澳大利亚健身补药 Vegamite 产生的"不可言传的"味觉，可以相当精密地言传为人们的四通道味觉系统中的"85/80/90/15 的频率调和"(一个最好避开的味觉空间死角)。而由一朵刚开放的玫瑰花所产生的"不可描述的"嗅觉，也可以相当精确地描述为人们嗅觉里面某个六维系统中的"95/35/10/80/60/55 的频率调和"。

这种更为深刻的概念框架甚至可以取代常识框架作为主体之间的描述载体和自发内省的载体。正像音乐家在将他所听到的音乐频率调和的内部结构的一般性理论内在化之后，就能学会识别这些音乐频率调和的构成一样，我们也能在将我们的主观感觉特性的内部结构的一般性理论内在化之后，

以内省方式学会识别这些感觉特性的 n 维构成。这种类比的更大的优点是取代这样一个可预测的响应：这种对“内部世界”的再构想会剥夺它的美丽和它的独特个性。这样做也许只不过和通过频率调和理论对音乐现象的再构想剥夺音乐的美丽和音乐的独特个性一样而已。与此相反的是，这种再构想打开了许多否则依然关闭着的美学之门。

我相信，这种常见的“难以用语言表达的感觉特性”所表现出的连续性带有一些显然确实可分成许多分量的特征。我们来看看用于面孔识别的人类“模件”。我们显然具有这样的东西，因为这种识别面孔的特定能力会因特定的右顶骨的损伤而破坏。这里似乎有理由提出一个也许是 20 维的内部状态空间表述，其中每一维对某个突出的面部特征如鼻子长度、面部宽度等进行编码。(警察的“容貌拼具”就是运用这种系统的尝试，已取得一些成功。)即使沿每个轴的辨别力只限于五个不同位置，像这样的高维空间仍将具有巨大的容量(5^{20} 个位置)，可以辨别和识别几十亿张不同面孔。它也能体现出相似关系，所以可以成功地对近亲加以分组，也能根据在不同年龄时拍摄的照片对同一个人加以确认。我们看一看爱因斯坦年轻时和年老时的两张照片。它们的相似之处在哪里呢？它们在人的面部状态空间中处在接近位置上。

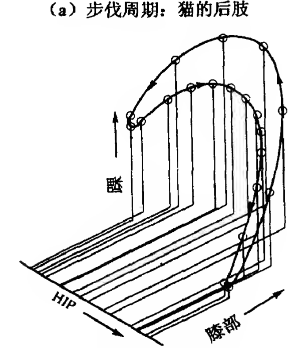
最后，我们来看看运动的例子，我们来考虑人的“身象”：一个人对他在空间中的整个身体构形的不断更新的感觉。这种构形是由数百块肌肉的同时存在的位置和张力构成的，人们从他们的大多数运动的平稳协调情况进行判断，对它的监控是相当成功的。人们是怎样做到这一点的呢？根据派利欧尼斯和利纳斯的理论，对于高维状态空间来说，是由小脑承担

了计算出实际的和意向的运动环境的高维编码之间的恰当变换的任务，这些编码是寄存在作为传入的平行纤维和作为传出的浦肯野神经轴突中的。

这里可以用一个很简单的例子再现一些这样的可能性。我们来看一个高度复杂的、周密编制的周期运动，如猫的运动器官活动情况(图14-12a)。现在考察猫后肢的三维关节角度



(a) 步伐周期：猫的后肢



(b) 状态空间轮廓图

图 14-12

运动状态空间，在这个空间中，该肢体的每一可能的构形都由一点来表示，而每个可能的运动由一个连续轨迹来表述。这

只飞跑的猫的优雅步伐的周期可很经济地由那个关节角度状态空间(图 14－12b)中的一个闭环来表示。如果以某种方式详细说明或“标记”出这个相关的环,那么协调运动这种令人生畏的任务就简化为一个简单的追踪问题:使你的运动状态空间位置沿着这个环的轨迹行进。

脑中是否存在着某种符合这一设想的东西,仍悬而未决。然而,目前曼尼托巴大学的拉里·乔丹 CNS 实验室正在进行这一技术的开发研究,他们的最终目的是把便携式微机用作截瘫患者产生有效运动器官活动的人工工具。

我们一旦跨越二维状态空间点的认知意义,进入了 n 维状态空间中的**直线**和**闭环**的认知意义,我们就有可能发现**曲面**、**超曲面**和**超曲面相交部分**等等的认知意义。出现在我们面前的将是一个对认知活动的不同于狭义句法概念的“几何”概念。

8. 结 束 语

我们已经看到,这种表述方案如何能以生物学上的真实方式对运动控制、感觉辨别和感觉运动协调的许多重要特征作出说明。但是,它有办法来说明那些所谓更高级的认知活动,例如通过**语言使用**以及一般地通过我们对世界的命题知识来表述的那些活动吗?

据我们设想,这是可能的。例如,人们可能设法找到一种表述“讲英语者的语言多维空间”的方法,这样,所有合语法的句子就会处于多维空间里的专门的超曲面之上,它们之间的

逻辑关系反映为某种空间关系。当然,我不知道如何做到这一点,但是它提供了一种可能性,可替代或是潜在地简化我们熟悉的乔姆斯基的描绘。

至于通常被认为构成个人知识的那种"信念集合",有可能是这种情况:语句的几何学表述使我们能够解决"默认信念"的棘手问题(Dennett 1975; Lycan 1985)。正像全息图不"包含"大量清晰的三维图像,它们以奇特的方式排列着,从而能在全息图被人从不同的位置观看时呈现出真实物体连续变化的景象一样,人类也很可能不"包含"大量清晰的信念,它们以奇特的方式排列着,从而聚集起来呈现出关于这个世界的一个连贯的说明。

更为可能的真实情况也许是,在这两种情况下,特定图像或信念只不过是一组更深的数据结构的任意投影或"切片",这种样本切片的总的连贯性是整体信息存储在更深层次上的方式的直接后果。例如,一台利用归纳规则来决定接受或拒绝单个得到的离散切片的归纳机器,会因过分谨慎细致而显得忙忙碌碌,但它不会得出总的连贯性的结果。这就意味着,我们要理解学习过程,可能还得理解直接支配着更深层次上的整体数据结构演变的作用力。

这些高度思辨性的评论显示了本文中概述的理论所提出的一个研究方向:一个通过坐标变换而相互作用的状态空间系统所具有的抽象的表述能力和计算能力究竟是什么?我们能够用它来系统地阐明代表"更高"形式认知活动的模型吗?这个理论也假定了相反方向的研究——脑的神经生理学研究的正确性。因为脑肯定不是一台以数字计算机方式工作的"通用"机,所以经常会出现这种情况:我们一旦事先准备好要

认清它们,就可以直接从脑的微结构中读取脑的局部计算策略。因此,关键是要研究这种微结构。(对认知神经生物学的通俗介绍,见 Churchland 1986。)

总的来说,通过坐标变换而相互作用的状态空间系统所具有的惊人的表述能力和计算能力,为理解神经系统的认知活动提供了一个强有力的、适用性极广的工具,尤为重要的是,适合于实现这种系统的物理机制遍布于整个脑之中。

参考书目

Allman, J. M., *et al.* (1982). 'Visual Topography and Function.' In C. N. Woolsey (ed.). *Cortical Sensory Organization*, Vol. 2, pp. 171–86. Clifton, NJ: Humana Press.

Arbib, M., and Amari, S. (1985). 'Sensorimotor Transformations in the Brain.' *Journal of Theoretical Biology* 112: 123–55.

Ballard, D. H. (1986). 'Cortical Connections and Parallel Processing: Structure and Function.' *Behavioral and Brain Sciences* 9 (1): 67–90.

Bartoshuk, L. M. (1978). 'Gustatory System.' In R. B. Masterton (ed.), *Handbook of Behavioral Neurobiology*, Vol. 1: *Sensory Integration*, pp. 503–67. New York: Plenum Press.

Churchland, P. M. (1979). *Scientific Realism and the Plasticity of Mind.* Cambridge: Cambridge University Press.

—— (1985). 'Reduction, Qualia, and the Direct Introspection of Brain States.' *J. Philosophy* 82 (1): 8–28.

—— (1986). 'Cognitive Neurobiology: A Computational Hypothesis for Laminar Cortex.' *Biology and Philosophy* 1 (1): 25–51.

Churchland, P. S. (1986). *Neurophilosophy: Toward a Unified Understanding of the Mind-Brain*. Cambridge, Mass.: MIT Press.

Cynader, M., and Berman, N. (1972). 'Receptive Field Organization of Monkey Superior Colliculus.' *Journal of Neurophysiology* 35: 187–201.

Dennett, D. C. (1975). 'Brain Writing and Mind Reading.' In K. Gunderson (ed.), *Minnesota Studies in the Philosophy of Science*, Vol. VII, pp. 403–15. Minneapolis: University of Minnesota Press.

Goldberg, M., and Robinson, D. L. (1978). 'Visual System: Superior Colliculus.' In R. Masterson (ed.), *Handbook of Behavioral Neurobiology*, Vol. 1, pp. 119–64. New York: Plenum Press.

Gordon, B. (1973). 'Receptive Fields in Deep Layers of Cat Superior Colliculus.' *Journal of Neurophysiology* 36: 157–78.

Huerta, M. F., and Harting, J. K. (1984). 'Connectional Organization of the Superior Colliculus.' *Trends in Neuroscience* 7 (8): 286–9.
Jackson, F. (1982). 'Epiphenomenal Qualia.' *Philosophical Quarterly* 32 (127): 127–36.
Kanaseki, T., and Sprague, J. M. (1974). 'Anatomical Organization of Pretectal Nuclei and Tectal Laminae in the Cat.' *Journal of Comparative Neurology* 158: 319–37.
Konishi, M. (1986). 'Centrally Synthesized Maps of Sensory Space.' *Trends in Neuroscience* 9 (4): 163–8.
Land, E. (1977). 'The Retinex Theory of Color Vision.' *Scientific American* (Dec.): 108–28.
Llinas, R. (1975). 'The Cortex of the Cerebellum.' *Scientific American* 232 (1): 56–71.
—— (1986). '"Mindness" as a Functional State of the Brain.' In C. Blakemore and S. Greenfield (eds.), *Mind and Matter*, pp. 339–60. Oxford: Blackwell.
Lycan, W. G. (1986). 'Tacit Belief.' In R. J. Bogdan (ed.), *Belief*, pp. 61–82. Oxford: Oxford University Press.
McIlwain, J. T. (1975). 'Visual Receptive Fields and their Images in the Superior Colliculus of the Cat.' *Journal of Neurophysiology*, 38: 219–30.
—— (1984). *Abstracts: Society for Neuroscience* 10 (Part I): 268.
Mays, L. E., and Sparks, D. L. (1980). 'Saccades are Spatially, Not Retinocentrically, Coded.' *Science* 208: 1163–5.
Meredith, M. A., and Stein, B. E. (1985). 'Descending Efferents from the Superior Colliculus Relay Integrated Multisensory Information.' *Science* 227 (4687): 657–9.
Merzenich, M., and Kaas, J. (1980). 'Principles of Organization of Sensory-Perceptual Systems in Mammals.' *Progress in Psychobiology and Physiological Psychology* 9: 1–42.
Mooney, R. D., *et al.* (1984). 'Dendrites of Deep Layer, Somatosensory Superior Collicular Neurons Extend into the Superficial Layer.' *Abstracts: Society for Neuroscience* 10 (Part I): 158.
Nagel, T. (1974). 'What Is It Like to Be a Bat?' *Philosophical Review* 83, (4): 435–50.
Pellionisz, A. (1984). 'Tensorial Aspects of the Multi-Dimensional Approach to the Vestibulo-Oculomotor Reflex.' In A. Berthoz and E. Melvill-Jones (eds.), *Reviews in Oculomotor Research.* New York: Elsevier.
—— and Llinas, R. (1979). 'Brain Modelling by Tensor Network Theory and Computer Simulation. The Cerebellum: Distributed Processor for Predictive Coordination.' *Neuroscience* 4: 323–48.
—— (1982). 'Space-Time Representation in the Brain: The Cerebellum as a Predictive Space-Time Metric Tensor.' *Neuroscience* 7 (12): 2949–70.
—— (1985). 'Tensor Network Theory of the Metaorganization of Functional Geometries in the Central Nervous System.' *Neuroscience* [16 (2): 245–74].
Pfaff, D. W. (ed.) (1985). *Taste, Olfaction, and the Central Nervous System.* New York: Rockefeller University Press.
Risset, J. C., and Wessel, D. L. (1982). 'Exporation of Timbre by Analysis and Synthesis.' In D. Deutsch (ed.), *The Psychology of Music*, pp. 26–58. New York: Academic Press.
Robinson, D. A. (1972). 'Eye Movement Evoked by Collicular Stimulation in the Alert Monkey.' *Vision Research* 12: 1795–1808.
Robinson, H. (1982). *Matter and Sense.* New York: Cambridge University Press.
Schiller, P., and Sandell, J. H. (1983). 'Interactions between Visually and Electrically Elicited Saccades before and after Superior Colliculus and Frontal Eye Field

Ablations in the Rhesus Monkey.' *Experimental Brain Research* 49: 381–92.
—— (1984). 'The Superior Colliculus and Visual Function.' In I. Darian-Smith (ed.), *Handbook of Physiology*, Vol. III, pp. 457–504.
Schiller. P., and Stryker, M. (1972). 'Single-unit Recording and Stimulation in Superior Colliculus of the Alert Rhesus Monkey.' *Journal of Neurophysiology* 35: 915–24.
Smith, D. V., *et al.* (1983). 'Coding of Taste Stimuli by Hamster Brain Stem Neurons.' *Journal of Neurophysiology* 50 (2): 541–58.
Stein, B. E. (1984). 'Development of the Superior Colliculus.' In W. M. Cowan (ed.), *Annual Review of Neuroscience* 7: 95–126.
Zeki. S. (1983). 'Colour Coding in the Cerebral Cortex: The Reaction of Cells in Monkey Visual Cortex to Wavelengths and Colours.' *Neuroscience* 9 (4): 741–65.



15 概念的联结论构造

A·屈森斯*

1. 前　　言[①]

解答认知体现问题的两个认知科学框架

认知科学理论是关于物理系统如何思维的理论。但是一个使认知科学理论化的框架,必须解释物理系统是如何**可能**思维的。意向性现象怎样才能成为自然科学所描述的同一世界的一部分?在这世界上怎么会有能够对这世界进行思维的有机体?这世界怎样能包含作为它自身一部分的对这世界的视角?我把由这些问题引出的可能性问题称为"认知体现问题"。

本文对认知体现问题所作的解答,具有心理学和计算两方面的特点。思维语言(LOT)框架(Fodor 1976,1987,并参阅§3)被列为一个候选解答,而作为其竞争者的认知科学框架("C3",即概念的联结论构造)也已形成。LOT和C3也都可当

做认知科学研究的方法论，它们有助于指导研究和理解研究的意义。因此，本文中将出现两种对照：关于认知科学事业的两种总的概念形成的对照，以及对于认知体现如何可能的两种理解方式的对照。

表述理论作为从计算模型得出心理学解释的工具

一个被看作具有心理学解释意义的计算用的人工制品就是一个“模型”。一个模型只是一个物理对象，怎样从中抽取出心理学解释呢？

认知科学理论（“一种理论”）以结构方式将基于模型功能的心理学解释结合为系统整体。这样，认知科学理论的建立就依赖于计算性的人工制品与心理学解释之间关系的概念的形成。而这种关系是以表述理论为中介的。

表述本身是一个具有双重特性的物理对象，即具有表述“载体”的特性和表述“内容”的特性。例如，标记外观上的一系列标记可以作为一种表述。这些标记所例示的字母数字混合符号的排序是表述载体的一种特性。如果这个序列碰巧如下：“斯坦福比牛津暖和”，那么该表述的内容就是斯坦福比牛

---

* 经作者允许，这篇文章在这一文集中首次发表，版权：1990。

A·屈森斯（Adrian Cussins），牛津大学哲学系研究员。

① 我通过注解使本文在一定程度上具有不同层次，以供背景不同的读者阅读。有一些注解是针对非哲学读者的，在这些地方我对文中用到的术语给出“简明的定义”。另一些注解是针对哲学读者的，我用注解指明与某个已有的哲学论题的联系。还有一些是一般性注解。

津暖和。表述载体是携带着作为其信息的表述内容的中介。

在一个模型中,表述载体的特性全部是具有计算效应的特性(例如,LISP码的句法特性)。它们是一些影响着模型发挥计算功能的特性。而构成表述内容的特性全部是具有心理效应的特性(例如,LISP码任务域的语义特性)。它们是一些影响着能从该模型得出的心理学解释的特性。所以一方面,表述的这些特性具有心理学解释的作用,而另一方面,它们又具有使模型发挥计算功能的作用。这种表述理论必须把这两组特性结合起来,从而建立起计算功能与心理学解释之间的联结。这一表述理论使我们能够从计算性物理对象中抽取出心理学含义,于是我们就得到了来自模型的理论。

如果认知科学包含从计算模型之中得到心理学理论,并且表述理论就是使之实现的方法,那么为了理解认知科学理论建立的实质,我们就需要理解计算、表述载体、表述内容和心理学解释之间的关系。这个任务从根本上说是多学科的(见图 15 - 1)。

(1) 心理学解释的种类→心理学,心理哲学
(2) 表述内容的种类→内容哲学,语义学 ┐
(3) 表述载体的种类→逻辑学,人工智能学,语言学 ┘ →表述理论
(4) 计算的种类→计算机科学,人工智能学

图 15 - 1　认知科学中的四个分析层次

认知科学框架

对图 15 - 1 四个层次中每一层次所作的分析,组成了认知科学框架。本文的主题是,这些分析并不是相互独立的。例如,如果已知对模型的计算构造体系所作的冯·诺伊曼分

析，以及对模型的表述载体所作的句法分析，那么该模型的内容理论就必须是语义学理论。而语义内容的选择势必要求从该模型中推导出一种特殊的心理学解释（概念论解释）（正如下文中解释的那样）。或者，如果人们选择了一个联结论计算构造体系，就可能被引向对句法表述载体的排斥，如斯莫伦斯基（Smolensky 1988 a）曾有过的情况一样。正如本文所表明的，这一结果本身将会牵连到联结论表述载体所能携带的各种内容，因而也牵连到能从联结论模型中抽取的各种心理学解释。每一层次的分析都制约着相邻层次的分析，因此也就可以自上至下地跟踪结果。

因而认知科学框架在各个这样的层次上都包含一个抉择，使得每一层次的抉择与所有其他层次的抉择都是相容的。对每一层次上的可能的选择的图示，提供了对竞争性的认知科学框架的表述。图 15－2 中那些表示选择的术语将在本文的正文中作出解释。但是以该图作为开始是不无益处的。

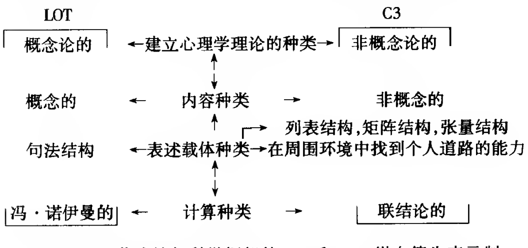


图 15－2 作为认知科学框架的 LOT 和 C3。纵向箭头表示制约。横向箭头表示每一层次上可能的种类或选择项目。图左边列出的是一组构成认知科学思维语言框架的选择项目，右边列出的是一组构成认知科学 C3 框架的选择项目。图中的陌生术语将在本文正文中加以解释。

本文的策略

根据 LOT 的预先假定，这里提出了一种可供选择的内容，其使用结果对于心理学解释、表述理论和计算实现方式来说，都是值得重视的。

如果可供选择的框架真的可以代替 LOT，那么它必须为认知体现问题提供解答；它必须指出认知的物理体现是如何可能的。第 2 节将解释为什么把该问题看作一个问题，并为它的解答提出必要和充分条件。

在 § 3 中，我将说明为什么认知科学的 LOT 解释为认知体现问题提供了一个候选的解答。我指出这种状况有赖于经典的**句法/语义表述理论（S/S 理论）**的计算应用。

建立 LOT 理论依赖于 S/S 表述理论，这势必要求基于 LOT 的心理学建模要运用**概念内容**。所以在 § 4 中，我说明了**概念内容**与**非概念内容**的区别，通过几个例子，指出了对非概念内容观念的心理需求，并引入了一种特殊的非概念内容：**理论构造内容**（CTC）。

在 § 5 中，我探讨了以概念内容建立模型的认知心理学结果：根据概念之间所具有的关系，以及感觉器/效应器与概念之间的关系，建立了心理现象的模型。这与 § 6 中提出的认为**非概念论的**心理学任务是解释**客观性的认知显现**的思想恰成对照。心理上的基本认知结构不是存在于概念**之间**的结构，而倒是概念**内部**的结构。第 7 节得出更为准确的**客观性**观念，并提供了一个评估任一系统作为概念运用系统的程度的方法。第 8 节建立了客观性与**视角无**

关性之间的联结。解释了为什么某些形式的非概念内容没有把世界作为客观世界呈现给主体，并解释了非概念内容得以客观地呈现世界的条件。第 9 节表明，转换非概念内容的心理计算理论怎样才能通过形成地图式认知结构降低系统能力的视角依赖性。这一点解释了：根据非概念内容建立模型的认知科学，怎样仍能满足对认知的概念制约。

我在结论中提出，联结论在认知方面令人感兴趣的应用，不应建立在 S/S 理论之上，因为 S/S 理论要求的是建立概念论理论，而联结论认知模型的建立适合于建立非概念论的心理模型。我提出了这种看法的理由：C3 适合于联结论（在认知上使用的方法），就像 LOT 适合于经典 AI 一样。联结论使用非概念内容的可能性，表明了福多尔和佩利舒（Fodor and Pylyshyn 1988）的批评为什么是无的放矢。联结论可以使用我介绍过的那些机构来说明：建立联结论认知模型在原理上怎样才能对认知体现问题作出回答。

2. 认知体现问题和对它的解答的构造制约

认知体现问题

我们来看看如下方式的认知体现问题的阐述。

根据本文的宗旨，我们假定存在一种对人类行为所作认

知解释的不可简化和必不可少的科学层次,同时假定,即使下一个千年期之末,认知科学也不会因神经生理学、量子力学或别的某个非认知解释层次而成为多余的。[1]

假定我们也承认自然主义:所有非物理特性或是可简化[2] 为物理特性,或是必须由物理特性来**实现**[3],或是必须由物理特性来**执行**[4]。换句话说,任何具有因果能力的事物,或是只具有物理的因果能力,或是必须由物理成分所构成,这样,原则上就可能理解,某个以物理方式构成的像**那个**(针对物理科学描述)的东西,为什么会具有**那些**因果能力(针对非物理描述)。自然主义并不要求非物理特性真的是物理特性,哪怕只是表面上的(自然主义不要求简化),但是它却要求:如果我们

---

① 根据本文的宗旨,我只是把这作为一个前提来假定。有充分的理由表明可以把认知解释看作不可简化和必不可少的。例如见布洛克所编书中福多尔一文(Block 1980: vol.i),福多尔的"计算和简化"(Fodor 1981 b),以及福多尔另一文(Fodor 1987, ch.1),佩利舒一文(Pylyshyn 1984: chs.1—2),帕特南一文(Putnam 1973)。我在另一篇文章(Cussins 1987 年 1 月)中论证了必然存在认知解释的科学层次。

② 只有当已被简化层次上的所有解释能够从作简化的层次上的解释推导而出时,才能使一个描述和解释层次简化为另一个层次(据 D. Charles)。如果一个层次上的所有特性与另一层次上的特性完全等同,那么可由此得出:这两个层次中的一个可简化为另一个;但是简化并不要求特性完全等同。重要的是要看到,我所赞成的(见后)认知的**构造**,并不需要我所不赞成的认知的**简化**。

③ 就消化是在胃中实现的意义而言。

④ x 实现 y,如果 x 是 y 的基底的话;但是关于什么是 y 的解释,是独立于(在任何方面都不依赖于)关于什么是 x 的解释的。认知科学的基本信条(遭到例如丘奇兰的反对,Churchland 1986)一直是神经生理学实现认知:认知事实可能由于神经生理学上的事实而成立,但是对认知事实的性质的解释是全然独立于对神经生理学事实的解释的。人们"仅仅"说到实现方式理论,并不是因为它提供的东西不重要,不合乎需要,而是因为它一点也不能阐明认知的心理本质。

了解所有能存在的科学，我们就不应认为特定的物理对象具有它们所具有的那些不可简化的非物理特性是出于偶然巧合。

无论人类行为是否能从生理学上得到解释，人类所表现出的行为方式是源自人类的神经生理学特性。但是根据我们的第一个假定，人类表现出来的行为方式是源自特定的不可简化的认知特性。神经生理学解释和认知解释是相互独立的，并且除非认知或生理出差错，它们在自身的说明方式上各自都是完备的[①]。那么以下两点怎样才能都为真呢：(1)行为的认知解释在因果上并非是多余的，(2)个人行为的物理起因与个人行为的认知起因是**齐步进行**的，所以一个人不会在个人的物理因果能力和认知因果能力的激烈竞争中被撕裂开来。我之所以如此写作，是因为我有如何尽量把哲学问题传达给部分非哲学读者群的想法，同样真实的是：我之所以表现出这种行为，是由于我身体中某些神经生理上的原因。我们怎样避免存在着一场控制我的手的斗争这一结论呢？

假定认知解释是非因果性的，那么这个问题没有得到解决，因为这样一来，这问题又会重新表现为：由物理方式引起的个人行为如何可能从认知角度看来是**连贯的**呢？我正在进行的写作，在认知上是可预测的(无论它是否由认知引起)，但是如果我的神经生理状况与数量浩大的可想象的方式中的任何一种方式有所不同的话，我就根本不会写下这些文字；因为

① 也就是说，除非有生理上的故障，用生理学说明方式表征的特殊标志效应，原则上可完全用生理学手段推导出来，而不必使用例如心理学定律或量子力学定律。这里也可能有一些例外，但它们没有普遍性。当然并不否认，特殊效应有可能在它的别的描述方式下以非生理学方式得到更令人满意的解释。用心理学说明方式表征的效应也有类似情况。

例如我的手可能不会动，或是固定在背后。我的生理状况怎样能持续地使我的身体去做限定范围的一些事情中它必须做的一件事情并使之具有认知意义呢?①

简言之，在既没有将认知特性简化为非认知特性，② 也没有将认知特性排除，③ 又没有否定认知特性的科学必要性④ 的情况下，我们怎样能从自然主义的角度理解认知呢?要理解这一点，就是要理解认知怎样能以物理方式体现，因而就是要理解怎样解决认知体现问题。

构造制约

对于简化、排除，以及解释的非必要性，存在着一种自然主义的抉择：由非认知特性**构造**认知特性。这一思想可从一个例子谈起。

有关建筑功能性的观念对建筑师的工作来说也许是最基本的，即使这个观念不能简化为施工人员配置建筑材料的观

---

① 我在一篇文章（Cussins 1987 年 1 月）中设计了一个思想实验，有助于清楚地认识这个问题。

② 认知简化论者有斯马特（Smart 1970），阿姆斯特朗（Amstrong 1968），普莱斯（Place 1970），刘易斯（Lewis 1966）。

③ 认知排除论者有奎因（Quine 1960）。P·M·丘奇兰（Churchland 1979），P·S·丘奇兰（Churchland 1986）和斯蒂克（Stich 1983）也经常被说成是排除论者，但是认为他们建议的仅仅是排除概念内容，而不是每个内容观念，也许是更恰当的。

④ 丹尼特（Dennett 1987）持有这种非必要性的立场：虽然存在着不可简化的认知特性，但是对于人类行为的完整的科学解释来说，它们不是必不可少的部分。严格地说，丹尼特认为不存在科学心理学，因为不存在自然的心理学种类（见 Cussins 1988）。S·希弗（Schiffer 1987）也是一位非必要性理论家。对希弗来说，存在着不可简化的认知特性，但只是在繁琐的修辞意义上。

念。例如,一个建筑师可能需要以一个有效的总体指挥部的观念来工作。但是这个观念不能根据砖、石、金属、玻璃、塑料、木材和水泥的商品尺寸的空间配置来定义。如果一个特定的公司处在特定的发展阶段上和特定的工艺及人种论背景下,那么对于一个有效的总体指挥部来说,由无数施工材料的不同配置构成的一个不可说明的无穷集将是充分的。不仅一个不可说明的无穷集使简化成为不可能,而且究竟是哪个不可说明的集合,也是随着背景参数而变化的。对一个小公司有效的东西,对 80 年代的 IBM 公司也许就是无效的。有了电话、电子邮件和传真机之后才有效的东西,如果放在早于这些通信媒介的工艺背景中,就会是无效的。

所以顾客向建筑师说明建筑物时使用的那些观念,不能简化为施工人员工作时必须用到的观念。因此存在两种不同的观念层次(描述层次),如果建筑师打算完成他的工作,他就必须以某种方式沟通这两个层次的观念。建筑师能够根据建筑学说明形成施工人员使用的说明,或是知道哪些建筑特性可从按照施工人员使用的说明建造起来的建筑物中得到例示,这种能力在外行看起来也许是难以理解的。并不存在建筑师使用的更深层的描述。事实上,在学习做这项工作时,建筑师已经获得了对建筑学观念和施工人员观念的理解,所以他可以在这两种层次上的描述之间来回行动。对建筑师来说,两种描述层次之间的关系是可理解的,①而不是偶然巧合,而对

---

① 我将"intelligible"(可理解的)一词的第一个字母大写,以指明它表示的是半技术观念。对这个观念的进一步讨论见拙作(Cussins 1990)。(原文此段中 intelligible 一词均作 Intelligible,作者在此说明了此用法的意义,但译文中没有反映这一点,请读者注意该词的特殊性。——译者)

外行来说，这种关系是不可理解的，所以看起来可能像偶然巧合。

建筑师的理解可能是实践多于理论。发现的在描述的建筑学层次与描述的材料层次之间的差距是可理解的差距，也许只要用下述技能：已知任何施工说明，一个建筑师就能说出[①]（并知道他能说出）以这种方式构造的建筑物会有何种建筑学特性，而已知任何建筑学说明（例如，提供一个与圣保罗大教堂背景相称的功能齐全的办公设施），一个建筑师就能知道如何把建筑材料组合起来，以满足这一说明。对熟练的建筑师来说，两种描述层次之间的差距是可理解的，而不是偶然巧合的：我们可以说，建筑师有能力从施工观念中**构造**建筑学观念，而外行不行。

因而，**构造**的关系就是层次之间的解释关系，这种关系不同于简化的关系，也不同于排除和非必要性关系。在一般应用中，**构造制约**要求任何非物理层次的描述和解释最终应该是能从某个物理层次构造出来的，这一点可类比于建筑师从施工和材料观念中构造出建筑学观念。如果我们必须从物理观念中构造出某个观念 Φ，那么我们必需能够根据一系列描述层次来理解对象是 Φ 这件事的性质，在这些层次中，最高层次使对象是 Φ 这件事显现出来（例如，建筑学层次使一个建筑物的总体有效性显现出来），而最低层次是描述的物理层次，同时每两个相邻层次间的差距是可理解的（而不是偶然巧合的），正如建筑师面对的建筑学层次和施工人员层次间的差

---

① 是感知到，不是做推理。推理关系从来只存在于一个层次中。感知（不管是不是感觉的）是跨层次的认知手段。

距一样。[①]

层次之间的差距或是**可理解的**，或是偶然巧合或不可思议的，[②] 根据我们对这两种差距间区别的直觉看法，构造制约可更为正式地叙述如下：

一个关于能力 Φ 的理论（或者更确切地说，是一个理论框架），当且仅当它在下述条件下根据拥有多于两个的一系列能力层次 L_1—L_n，对一个有机体拥有这种能力的事实作出解释时，满足对 Φ 而言的构造制约：

1．基础层次 L_1 的能力使我们不理解为什么具有这些能力的有机体就是具有能力 Φ 的有机体（即，L_1 和 L_n 的那种联系，会产生出关于它们如何齐步进行的不可思议的偶然巧合问题）；

2．这理论说明为什么拥有最高层次 L_n 的能力就实现或显现了拥有 Φ；

3．对每一对层次 L_i 和 L_{i+1}（$1 \leqslant i < -$）来说，L_i 和 L_{i+1} 的联系方式不会使它们产生出关于它们如何齐步进行的不可思

---

① 两个层次之间的可理解的联结不包含层次之间的定律，也不包含借以理解两个主要层次之间联结的第三个描述层次。这就是我强调建筑师能力的实践特点的原因。如果两个层次的齐步进行并不表现为不可思议的偶然巧合，这两个层次之间的联结就是可理解的（Cussins 1987 年 1 月）。其中一个方面是，某个掌握了可理解联结的人，应具有一种可出错的、实践的能力，以安排较低层次的系统，使之满足较高层次的制约，同时应具有某种能使较高层次性能发生蜕变的环境的观念。我看不出有什么充足的理由要使可理解的差距与不可思议的差距之间仅属直觉的区别（它取决于例如我们把什么识别为可理解的）成为一个问题。构造制约是对解释作出的制约。情况可能是这样的：什么东西要成为解释性的，本身只能根据人类发觉是解释性的东西——对人类来说它是一种解释——来解释。如果能够以不同于此的方式说明可理解的差距相当于什么，那当然很好，但是就我的目的而言，我们能够说明层次之间的任何差距已经足够，不管它是可理解的还是不可思议的。像塔尔斯基的约定 T 一样，构造制约是关于成功的判据，我们能够识别是否满足这一判据。并且像塔尔斯基的约定一样，它依赖于一个未定义的观念。（是 D·查尔斯使我注意到这一对应关系。）

② 我在另一篇文章（Cussins 1987 年 1 月）中，给出了一个关于不可思议的偶然巧合的思想实验。

议的偶然巧合问题(即 L_i 与 L_{i+1}之间的差距是可理解的)。[①]

习俗心理学层次与神经生理学层次之间的差距是如此地不可理解:我利用这种差距,使得作为控制我的手的一场激烈斗争的这一认知体现问题变得鲜明起来。如果仅仅给出习俗心理学和神经生理学知识,这两个层次的齐步进行看起来是一种不可思议的偶然巧合。[②] 与之不同的是,因为我们知道如何由电子元件建立机器语言,以及由机器语言建立像 LISP 那样的高级语言,所以计算机的行为取决于两个各自在自己的说明方式下是完备的描述(LISP 和电子)层次,这一事实看来并不是偶然巧合。LISP 与电子之间的构造关系,并不要求在任何两个相邻层次之间存在着定律式的关系:从未听说过联结电子描述和计算描述的规律。所要求的仅仅是,某个具有例如机器语言、LISP 语言和电子学的知识的人,应当大体上知道如何组装电子器件,以造成一台使用 LISP 语言的机器。并不是说一个人每次都必须成功,而只是说一个人不应

---

① 心灵哲学家们感兴趣地注意到,构造制约在某一方面弱于并发现象(甚至是整个物理世界中的并发现象,而不仅仅是头颅中的),在另一方面又强于并发现象。说它较弱,是因为它不要求一个构造中的较低层次对于较高层次必须是充分的,像先发层次对并发层次必须是充分的那样。所以联系到 §2,我们可以注意到,与 LOT 相反,句法并不决定语义,因为在给定相同的句法和证明论的情况下,总是存在对于例如连接词的不同的语义解释(Williams 未出版)。尽管如此,在任何可能的语义解释之下,句法仍然保持语义制约,这一事实也许足以说明运用语义观念对行为的解释(心理学解释)与依靠句法的计算实现方式对行为的解释齐步进行的原因。因而,虽然句法不决定语义,但这个事实不能排除在认知构造中使用句法/语义表述理论。

说构造制约比并发现象强,是因为它对解释提出了并发现象没有提出的要求(并发现象不要求层次间必须有可理解的联结)。

在我看来,构造制约与并发现象相比的一个优点是:我知道如何论证构造制约,但我不知道如何论证并发现象。

② 对某些作者例如斯温伯恩(Swinburne 1986)来说,齐步进行不仅看起来是,而且就是不可思议的:为了解释一个人行为的连贯性,必须祈求于上帝。

感到他制成的东西具有计算机的功能是偶然巧合。

有一种对认知科学的解释,按照这种解释,认知科学一直在探索概念能力的构造[①],采用的方法是在习俗心理学与神经生理学层次之间插入表述层次和计算层次。在下节里,我将指出怎样可把 LOT 理论视作提供了一个沿着这些思路的说明。文章的其余部分提出了另一种说明,更适合于联结论计算构造体系的应用,并且不会遇到像 LOT 遇到的同类困难。

3. 作为概念构造的思维语言

值得注意的是,通过满足对拥有概念的构造制约,来解决认知体现问题,在这方面只做过一次认真的尝试。[②] 这就是福多尔称为思维语言(LOT)的模型,他和其他人曾在哲学层次上为它辩护了十多年。LOT 可能与否,关键在于从弗雷格以来已在逻辑中得到发展的表述理论在计算和心理两方面应用的可能与否。正是这一理论,根据它的组合句法和组合语义,表征了一个表述系统。

---

① 对于构造制约中用到的术语"构造",我一直是从技术的意义上使用的。两个相邻层次间的关系可以是一种构造关系,即使构造制约要求多于两个层次。如果它是可理解的,它就将是构造关系。对理论框架的自然主义制约,并不是要使每两个层次间的关系成为构造关系,而是要满足构造制约。在一个满足构造制约的框架中,非相邻层次间的关系也许不是可理解的关系,因为对它的掌握可能取决于对中介层次理论的掌握。

② 构造理论在开始阶段是以信息的通信理论观念为基础的(Dretske 1981)。和行为主义一样,这些理论的困境是,虽然信息观念适合于构造中的低层次,但是没有一个人哪怕是在原则上证明了如何可能在信息层次之上构造一个概念拥有层次。德雷斯克在他书中第七章这样试过,但遗憾的是这个尝试失败了(Cussins 即将出版)。提出别的非简化主义和非排除主义的理论框架的有米利肯(Millikan 1984),但是还不清楚这个框架怎样能产生一个构造。

表述系统的句法理论为这系统的所有一串串合法的原子表述提供了一个递归式说明。一个表述系统的语义理论为所有合法表述的解释提供了一个可公理化的、递归的详细说明。如果这一表述系统要起到什么作用的话，就必须有可能在句法之上，或是定义一个用于逻辑作业的证明论，或是定义一个用于计算作业的过程结果理论，这个理论根据每一合法表述对所有合法变换作出了说明。逻辑传统的价值大多建立在我们能够用纯句法形式定义合法变换理论的基础之上，尽管它与语义制约有关。

句法和语义理论在解释上必须是独立的，然而又是有联系的。它们必须是独立的，因为在一点也不了解语义理论的情况下，必须有可能理解如何应用合法变换理论；但是它们必须是有联系的，因为合法变换理论的句法应用不可违反语义制约：按照惯例，它不可把一组正确的前提变换成一个错误的结论。同时，句法理论和语义理论必须相互联系得即使表述系统不完备也必然能通过合法变换理论的句法应用获得所有语义连贯的变换中的有用部分。

LOT之所以如此引人注目，是因为它深刻地说明了：获取认知科学建模的计算组成部分与心理组成部分之间必要的可理解的联结方法，是通过形成一个句法和语义表述系统，对这系统来说，句法是以计算方式实现的，而语义[①] 则适合于心

---

① 对LOT的狂热支持者来说：LOT里的很多论点都围绕这一点在转：心理学解释是否需要一个特殊的语义学观念。这一特殊观念一般被称为“狭义语义学”，经心理学解释后，它变为“狭义内容”（Fodor 1987）。狭义内容并不完全确定参照物。狭义内容可能必须是一般的，而内容的所有单独方面则被当做广泛的现象，同时狭义内容也可以限于表现观察特性，而排除自然类特性。但问题是这些新方法受到很强的制约：狭义内容和狭义语义学必须是狭义的——它们必须构成经典内容和经典语义学的一个子集合。

理学解释。如果能做到这一点,那么取得计算与心理学之间的必要关系的方法,就可以直接从我们对如何建立表述系统的理解中得出,因为该系统有可能说明一个与语义制约有关的合法变换的句法理论。S/S 理论给出的是一个与语义层次齐步进行的、独立于语义的(句法)层次。所以如果句法能由计算实现,而语义学能提供心理学解释的基础,那么 LOT 就已经说明了计算转变怎样能与心理学转变齐步进行。

经典计算构造体系的句法表征方式是很自然的,其部分原因是这些构造体系本身是从逻辑传统发展而来的。语义学的心理学应用也是自然的,这是由于可追溯到弗雷格的一个传统;这个传统就是把一个句子的意义看做命题态度的对象,它是由心理学动词("相信","愿望",……)、"那个"和句子本身串连在一起来表达的。

正如 LOT 所解释的,认知科学变得与计算理论和心理学理论两者都有所不同,因为它试图以经验方式确立一个建造在认知表述系统之上的中间解释层次,使得:(1)句法在计算上是可实现的,(2)语义获取了重要的心理概括,(3)过程推断理论是一致的,并且使用起来是完备的。实际情况当然是,如果没有 LOT 认知科学研究也将继续前进,很多认知科学家不赞成 LOT,并且提出了很多反对 LOT 的意见。但是 LOT 仍然是唯一提供了评价认知科学成功与否的明确判据的理论,它似乎有理由是可行的,它还表明认知科学怎样可能起到它希望起的重要作用:从计算的组成部分出发构造认知科学模型的心理组成部分,从而解释怎样能以物理方式体现认知。

我这里的目的不是要评论 LOT,而是要理解如何形成一

种可供选择的、具有可比较的解释范围的理论。LOT之所以给人以深刻的印象，是因为它建立在S/S表述理论的引人注目的传统之上。要就认知体现问题形成一个可供选择的解答，我们必须形成一种可供选择的表述理论。一个可供选择的**计算**理论是不够的。

4. 内容、概念内容和非概念内容

我已经指出，就LOT而言，提供适合于认知体现问题的那种解释，怎样取决于LOT所使用的表述理论。但是，表述理论的选择决定着一个模型所能提供的那种心理学解释，这也是事实。所以如此的原因，是表述理论决定着能被指派给模型状态的那种**内容**，而这又决定着模型能使其成为可利用的那种心理学解释。表述与心理学解释之间的联系就是**内容**。

内容简介

在作出更细致的分析之前，我先就内容、概念内容和非概念内容的观念作一简要说明。

人类按他们的方式行动，从而人类也常常按他们的方式规范行止，因为世界的某个方面是以某种方式向人类呈现的。"内容"这一术语，正如我要使用的那样，在第一种情况下是指世界的某个方面呈现于主体时所采用的方式；客体、特性或事物状态就是以这种方式在经验中或思想中给出，或是向经验

或思想呈现的。例如，我把我面前那个灰色的塑料长方形物体看成是一个打字键盘，它带有常见的 Qwerty 结构。我也看到它**存在于**我面前，而这些事实成为我的手以一定方式来回移动的原因。我的表述状态具有内容，借助于这内容，表述状态使我能够理解世界，从而指导我的行动，并且它们（通常）作为正确的或不正确的事物呈现给我。我要谈一谈**具有内容的表述状态（或载体）**。单个的表述载体可能带有不止一个内容，甚至不止一种内容。

内容理论——我们据以解释内容是什么——确定观念相对于我们的经验、思想和世界的观念的位置。但重要的是要看到，这与被应用于并不是经验主体状态的一些状态（虽然不利用这些状态来解释）的内容观念是一致的。[①] 这一观念存在着一些派生用法，应用于起交流作用的认知产物，诸如说话、书写和别的信号系统，或是应用于人的非意识状态，像下意识信息加工状态，但这些用法最终必须根据认知经验中内容的初等应用理论来解释。[②]

概念内容是指这样的内容：它呈现给主体的世界是客观的、具有人类特征的世界，关于这个世界，人们既可能作出正确判断，也可能作出错误判断。如果存在其他各种**非概念**内

① 甚至假定：在任何经验主体存在之前就有各种简单的内容在世界各处存在着，这也与此相一致——关于这一点，以后详述。

② 在这种背景下，这简直就是规定。那些以非弗雷格语义学传统，而不是达米特/斯特劳森传统从事研究的人，会感到我的用法有些奇特，所以我一开始就提出了这个规定。我需要有一个像我的内容观念一样的观念——无论把它称做什么，因为认知体现问题部分地是要解释：特定的、像人类那样而不是草履虫那样的自然界生物怎样才能存在，这些生物对世界的响应并不完全是它们对感觉外表的物理刺激的响应，而是部分地取决于关于这一世界情况的概念。

容，那么它们之所以存在，是因为有一些方式，通过它们，即使客观的、具有人类特征的世界不为主体所理解，也可以将世界呈现给经验主体。假定必然存在着非概念的内容形式，并非是不合理的，因为我们关于人类幼小的婴儿（比如说在获得客体概念之前），或是耄耋之年的老人，或是某些别的动物，所要说的就是这种事情。认为这些生命体具有经验的想法是完全可取的，然而他们不能与我们交流思想；我们不能（从内部）理解他们是如何对世界作出响应的；我们也不能把我们的世界强加于他们。

概念内容把世界呈现给主体时，将世界划分为对象、特性和情境：真值条件的组成部分。例如，我有一个复杂的概念内容（思想）：今天老城墙笼罩在雾中，这一思想呈现给我的世界的样子，就是今天老城墙笼罩在雾中这一事物状态是存在的。为了理解这个内容，我必须认为世界是由如下方面组成的：对象——老城墙，特性——笼罩在雾中，并且前者满足后者。拥有任何内容都免不了以这样或那样的方式去剖分世界。如果经验提供了一种剖分世界的方法，不是把世界剖分为对象、特性或情境（即真值条件的组成部分），就会存在一种**非概念**内容的观念。[①]

拥有内容就是**形成**关于世界是如此这般的**概念**，这是一

---

① 怎么会有像这样的方式呢？这正是我用这篇文章中很多篇幅试图解释的事情。在这一节里，到现在为止，我认为我已经给出了对内容、概念的和非概念的这些观念的简要说明。这一节的其余部分将开始分析，并论证非概念内容的存在，而在 §7 中，我将更准确地论及这一观念：内容客观地将世界呈现为由对象、特性和情境组成。本节的论断没有告诉我们内容是什么，只是用来给出对于这些观念的直觉感受。下面我们将看到，概念内容可在任务域中由经验获得，而非概念内容可在基础域能力中由经验获得。

种很自然的说法。但是“概念形成(conception)”这个词与“概念(concept)”的关系太密切了,以致它不能在世界的概念表现与非概念表现之间中立地发挥作用。我要说的是,[①]内容把世界**记录**为某种存在方式,那么请问,是不是存在着不把世界记录为对象、特性和情境的记录世界的方式呢?

概念特性和非概念特性的定义

我打算先通过引入概念和非概念**特性**的定义,更仔细地分析这些观念,然后说明这些定义怎样才能在内容理论中得到应用。[②]

一个特性是一个**概念特性**,当且仅当相对于理论来说,它只通过概念典范地被表征,[③] 而这些概念是有机体为了满足这一特性所**必须具有**的。

一个特性是一个**非概念特性**,当且仅当相对于理论来说,它通过概念典范地被表征,而这些概念是有机体为了满足这一特性所**不必具有**的。

请注意,这两个定义的区别主要在于第一个定义里所强调的“必须具有”与第二个定义里所强调的“不必具有”之间的区别。

我们来看一看认为某人是单身汉这一想法的特性。要说

---

① 根据贝内特(Bennett 1976)和B·史密斯(B.Smith 1987)的用法,但稍有不同。

② 注意:这些并不是两种内容的定义。

③ 某一事物(在一个理论内部)典范地被表征,当且仅当它通过该理论认为是它的必要特性的那些特性被表征时。例如,在足球协会里,足球比赛的典范表征所依据的是比赛规则中使用的观念,而不是运动场受到破坏的短暂模式。一个内容的典范表征所靠的是使它呈现世界的方式得到揭示的一种说明。见后。

明这一特性是什么，将用到*男性*、[1] *成年人*和*未婚*这些概念。除非拥有[2]这些概念，否则就不能满足这一特性，因为除非他或她能够认为一个人是男人、成年人和未婚的，否则就不能算做是认为这人是单身汉这一想法。所以认为某人是单身汉这一想法的特性(与是单身汉这件事的特性不同)是一个概念特性。

我们也可以看一看这个信念特性：相信斯坦福大学校园在这儿附近(这里我把斯坦福大学校园认作斯坦福大学校园，而不是西部最富有的大学的校园，并且我把这儿认作这儿，而不是狼山路 3333 号)。如果已知这一点，那么除非拥有作为斯坦福大学校园的这一斯坦福大学校园的概念，否则就不能满足这一特性。因而这一特性的典范表征，只是通过有机体为了满足这一特性所必须具有的概念，所以这个特性是一个概念特性。我们再对比一下具有一个活动的下丘脑这件事的特性。这样一个特性是通过*下丘脑*这个概念来表征的，但是一个有机体在不拥有这个概念的情况下也可以满足这一特性。所以具有活动的下丘脑这件事的特性是一个非概念特性。[3]

从形式上看，这一思想就是，概念内容是由概念特性组成的内容，而非概念内容是由非概念特性组成的内容。对这一形式思想，我们能提供什么实质性的东西吗？

---

① 我像使用引号一样使用星号，是指星号内的词并不具有它们通常的意义。但是星号是指这些词代表的是这些词所表达的一个概念或一些概念，或是另一种内容，而不是语言条目本身。

② 注意例示、满足或隶属于一个概念与拥有一个概念之间的区别。我拥有*单身汉*这一概念，但我不隶属于这个概念。

③ 显然不是一个内容特性。

概念特性和非概念特性定义在内容理论中的应用

为了说明存在着非概念内容的观念，我们需要说明在内容理论中能够应用非概念特性的定义。这意味着什么呢？

概念特性和非概念特性的定义，所用的是典范说明的观念，否则每个特性都会成为非概念特性，因为一般说来，每个特性——包括概念特性在内——都可以通过主体所不必拥有的概念来说明。所以我们需要运用**典范说明**的观念。如果我们打算把这些定义应用到内容理论中，那么我们感兴趣的典范性的观念就是成为内容理论中的典范说明这件事的观念。当一个状态或一个活动的某种说明是内容理论为了获得世界的某些方面而被给予状态或活动的主体时所用的特殊方式而产生的说明时，这种说明在内容理论中被确认为是典范的。所以正像麦克道尔(McDowell 1977)阐述的那样，"'aphla'代表aphla"是典范的，而"'aphla'代表'ateb'"则不是典范的，虽然两者都是正确的，因为aphla就是ateb。在内容理论中是典范的这件事的观念与在数字理论中是典范的这件事的观念有类似之处，对后者而言，数字9的典范说明不是"行星的个数"，而是"9"。

概念内容的情形

任务域观念。为了了解概念内容是如何工作的，我们需要一个关于行为的任务域观念。**任务域是世界中的一个有**

界区域，它被看作已经记录在一组对象、特性或情境的已知组织形式之中，[①] 它不包含任一个或任一些有优先权的观点，行为的评价是就它而言作出的。[②]

SHRDLU[③] 的积木微世界是 SHRDLU 的任务域。形式语义学中模型的观念，以及(往往)逻辑学中可能世界的观念，都是任务域观念。同样，国际象棋计算机性能的评价是就国际象棋任务域而言的，组成该任务域的有以下方面：分成两类的 64 个方格，32 个子——每个子都有自己的身份特性，一个法定的开局位置，三种法定的终局位置，和一组从每个合法位置到从这些位置作出的所有合法延续的变换。这种计算机的任务域排除了诸如人的情绪和计划、照明条件、获胜的理由、获胜的要点等因素。也就是说，对国际象棋计算机性能的评价是就 64 方格盘上国际象棋标志的变换而言的，而不是就它对人的情绪、照明条件、历史上的棋局模式或者"它获胜的理由"所作的响应而言的。此外，因为这一领域是固定的，使得某些情境记为白方胜，而另一些记为黑方胜，所以计算机性能的评估不是就它把它的知识转换到一种不同的游戏——象棋*的能力而言的，除非象棋

---

① 任务域的对象、特性和情境被假定为完全客观的，就是说假定在原则上有可能用一个独立于任何一种对有机体的识别、感知或作用于它们是怎么回事的解释方式来解释它们的存在是怎么回事。(由此表明，任务域观念是一种理想化形式。)重要的是要看到任务域完全是从感知者或主体中抽象出来的。任务域中没有观点，也没有基本上是指示性的元素。

② 人们可以假定一个任务域就是世界的一部分。但情况并非如此，因为**在一个给定概念化的条件下**一个任务域才是世界的一部分。世界不仅允许有许多不同的正确的概念化形式，而且也允许作为非概念化形式的一些记录(我将作出论证)。

③ 见威诺格拉德(Winograd 1973)。

中白方胜的那些情境在象棋*中是黑方胜，而在象棋中黑方胜的那些情境在象棋*中是白方胜，否则象棋*完全等同于象棋。[①]

由此可知，任务域就是世界的一个概念化区域，它提供了评价某个系统的性能的背景（真/假，赢/输，在模型中为真/在模型中为假，适应/不适应，成功/不成功……）。任务域观念是如何与概念内容的观念相联结的呢？

用任务域概念说明 α 内容。再来看看我身上发生的认知事件，我们用话语把它表达为："我在想斯坦福大学校园在这儿附近。"这是我的一个表述状态，可能拥有不止一种内容。[②]这个状态带有哪种内容呢？有一种内容的类型（我们称它为"α内容"），它在内容理论[③]中被规定为是一种具有确定的[④]真值

---

① 为了使这样的游戏得以进行，必须补充新的规则，例如强行吃子规则：轮到走子时，如果一个棋手能吃对方的子，那么他必须去吃。但这并不改变这一点：下棋的智力能力（不同于常规计算机的能力）必须使这种能力也变得能下相关的棋类，这些棋类的任务域可能互不相同，并且也不同于原来棋类的任务域。

② 我们不应假定：由于该状态具有语言学表达方式，所以它只具有一种内容，因为语言学内容并不是一种内容，而只是内容的一种**表达方式**，或载体。事实上，我们需要不止一种内容才能正确对待我们的语言使用。在本文的当前阶段，我打算在这一点上持中立态度。

③ 例如见达米特（Dummett 1975 and 1976），戴维森（Davidson 1967），埃文斯（Evans 1982：chs. 1—4）。

④ 具有概率真值条件是具有确定的真值条件的一种方式。当奎因论证"gavagai"的参照物不确定时（Quine 1960），他的意思并不是它以某一概率表示兔子，以某一概率表示兔子生长阶段，以某一概率表示联结着的兔子部位。根据我的看法，模糊集合论和语义学理论的概率性校订并没有向我们提供不同于概念内容的内容观念。相反，它们提供了一种方法，根据这种方法，一个状态或一个条目等可以以概率方式具有概念内容。在一个任务域中抛掷的硬币，正面朝上的概率可能是1/2。任务域是完全确定的，但不是决定性的。

条件的内容；[1]即它的评价作为一个正确评价，把确定的条件施加于世界。因此，“斯坦福大学校园在这儿附近”这一语言表达，不能完全获得这一表述状态的 α 内容，因为这需要对“附近”和“这儿”作出固定的解释。（为了使这一状态成为带有 α 内容的状态，我们需要知道它施加于世界的是何种真值条件。但是“这儿”和“附近”这两个词没有告诉我们这一点。）

现在假定这种状态是作为我的规划的一部分出现的，在这规划中，我正在计划如何在已知各种参数和对我的制约即时间、财力、饥饿程度、吃饭地点的远近、我所能利用的运送速度、各个地点的食物价格的情况下最满意地吃午饭。这些参数和制约建立起一个任务域，它固定了对术语“附近”和“这儿”的解释：假定根据我的时间制约和饥饿程度得知我需要在15分钟内吃上午饭。那么“附近”意指：通过我所能利用的运送方式，能在15分钟内抵达。同样，“这儿”将意指某种像这样的事物：在我站立的地方与所有属于我的计划域的运送模式的出发点的连接线之间的区域。

将我的认知发生状况解释为具有 α 内容，取决于通过任务域概念对该内容作出说明；在这情形中，则取决于各种制约和各种给定参数下我计划吃午饭的任务域。换句话说，为我的认知状态提供确定的真值条件，尽管这是把它解释为具有 α 内容所需要的，却势必要求这一内容通过反映该任务域的客观结构即它的表现为对象、特性和情境的组织形式的一些

---

① 对于是否存在着不止一种内容的争论，我不抱什么偏见。我注意着内容理论内部的某种制约，并把满足这一制约的内容称做“α 内容”。稍后，我将介绍内容理论内部的一种不同的制约，并把满足这一新制约的内容称做“β 内容”。至于“β 内容”也许等价于“α 内容”的问题，则悬而未决。

概念而得到典范说明。因为一个有机体只有在掌握它的真值条件(或者它对包含它在内的内容的真值条件的贡献)时才能掌握一个α内容,所以掌握这种内容的有机体必定知道内容的任务域(的相关部分)是什么。但是任务域(不同于世界)基本上是以概念方式构成的,所以如果不拥有据以构成任务域的概念,就无法知道内容的任务域是什么。因此,要拥有一个α内容,就必须拥有据以对它作典范说明的那些概念。由此可见,α内容正如上面定义的那样,是一种由概念特性组成的内容。即,α内容是概念内容。

要证明非概念内容的观念,也可以仿照确认α内容为概念内容的过程。我们必然要问,在内容理论中以类似的方式促成非概念特性定义的应用的方法存在吗?在提出这一问题时,我是在问,状态或活动的非概念说明是否能在内容理论中成为典范,因而我也是在问,要获得一个把世界的某个方面给予活动的主体的特殊方式,一个正确的内容理论是否能**要求**对该活动的非概念说明。

为了弄清楚这样做时包含些什么,我们可以列出在促成内容理论中概念特性的定义时我曾使用过的不同要素,作为对上述讨论的总结:

1. **概念特性**的定义(由规定得出);

2. 论断:在**内容**理论中存在着一个要求有确定的真值条件的制约,[①] 同时,拥有满足这一制约的内容,要求有关于真值条件的知识(掌握真值条件)(这些论断是由内容理论给出

① 或是对确定的真值条件的确定贡献。我将不再继续采用这种限定性的说明。

的，并且是构成这个内容观念的）；

3. 以语言方式表达为“想着[①] 斯坦福大学校园在这儿附近”的心理状态，这状态还没有就它所具有的那种内容作过分析；

4. 论断（本文已作过论证）：在（2）的条件下对（3）作出解释，要求的是任务域观念，以及通过任务域概念对（3）的内容作出的说明。

5. （4）导致了对（1）的满足，从而确认满足（2）中制约的内容是**概念**内容。

任务域的观念提供了α内容的哲学观念与我对概念特性的规定性定义之间的联系；为了证明要根据α内容作出心理状态的分析就必须满足概念特性的定义，这一联系是必不可少的。

非概念内容的情形

我能够证明对非概念内容的需求，其方法是证明存在着一些心理状态，要充分理解它们，要求有不能用上述方法进行分析的内容观念；即这些心理状态必须通过主体无需具有

---

① 从弗雷格以来，“想”是一个心理学概念，而“思想”是一个逻辑或哲学观念，这已成为哲学上的常规。像“思想”一样，“概念”首先是一个逻辑观念：概念是思想的结构成分。所以在这一常规内，“我关于P的思想”的说法要求这个状态的内容必须是概念的。而仅仅说“我想着P”则不要求必须有关于这个状态所具有的那种内容的任何结果。

我在本文中所说的内容的一部分是这一问题：“概念”在作为逻辑观念的同时，是否也应当是心理观念。我假定，心理学必须运用某种内容观念，但我要指出，那种作为概念内容的内容只具有逻辑-哲学的作用：心理学要求有一种不同的内容：非概念内容。

的概念,才能典范地加以说明。这一讨论应当对照概念内容情形的讨论进行,所以我们需要列出与(1)—(5)对应的要素如下:

1′. 非概念特性的定义(由规定得出);

2′. 对一种内容即β内容提出的某些构成条件,它们是由内容理论提供的,但是不同于(2)中的条件;

3′. 某种心理状态或表述状态,尚未就它所具有那种内容作过分析;

4′. 论证下述论断:在(2)′的条件下对(3)′作出解释,要求的是任务域以外某个域的观念,以及通过该域的概念对(3)′的内容作出的说明;

5′. 证明:(4)′导致了对(1)′的满足,从而确认β内容是非概念内容。

我们已经有了(1)′。(2)′的情况怎样呢?

认知意义。一个好的内容理论是与各种制约相应的。例如,一个好的内容理论应该适合于在一个基于内容的科学心理学中使用,它应该有办法解释某些内容怎样具有确定的真值条件,并且一个好的内容理论也应该获得**认知意义**,即这一内容在感知、判断和行动方面所起的作用。

内容理论如何能为认知意义提供方便呢?为了解释某些同一性陈述如何可能是含有信息的,首先引入弗雷格的意义观念[①]。例如,了解到"长庚星 = 长庚星",并没有学到任何新的东西,而了解到"长庚星 = 启明星",有可能学到某个有重要意义的东西,因为长庚星的确就是启明星。由此可见,拥有这

① 弗雷格(Frege 1891)。

里由"启明星"一词所表达的内容，不可能仅仅在于想起金星这个行星（没有比这更深一层的说明）的能力，因为同样的能力也是与"长庚星"连在一起的。这里存在着一个促成引入内容（意义）观念的因素，这种观念不同于纯参照的内容（参照物）观念。存在一个由"长庚星"表达的内容，它不同于由"启明星"表达的内容，因为在个人对具有"……=长庚星"形式的内容的真值判断中，前一内容起着不同于后一内容的作用。弗雷格把这种促成概括为这些内容（意义）的同一性判据。[①] 我们还可以对它作进一步的概括，从而得出一个一般化的意义观念，我称它为"β 内容"，它的同一性条件是固定的，这并非仅靠它与判断的构成性联结，而是靠它与感知、行动和判断的构成性联结。[②] 拥有特殊的 β 内容，要求拥有在主体的心理体系中起这作用的内容状态，该体系是构成 β 内容的。

内容理论近期研究中取得的一个重要进展，是证明存在着指示性和示范性的 β 内容，该内容不能以合乎概念内容的方式，通过任何描述作出典范的说明。[③] 取得这一认识的方法，是证明，要是存在着一种——**照惯例是不可能的**——描述，可用合乎概念内容的方式，为该内容提供典范说明，那么这种描述将会改变内容的认知意义，即改变它与行动和判断的构成性联结的性质。由于认知意义是构成 β 内容的，因此这种说明形式不能以典范方式获得 β 内容。

---

① 见本书第 524 页注①。

② 例如，见皮科克（Peacocke 1986）。弗雷格对意义增加了进一步的条件，即它确定参照物。然而这不是对 β 内容的条件。只有某些 β 内容（它们是意义）才确定参照物。

③ 当然，我提到的那些人并不这样提出他们的结论！

例如，佩里(Perry 1979)就指示性的"我"以及与行动的联结证明了这点，而皮科克(Peacocke 1986)就示范性感知内容以及与感知和判断的联结证明了它。佩里的观点是，对于指示性内容*我*，在概念上使用任何描述性的典范说明——*x，所以 Φx*，将通过改变它与行动的构成性联结，而改变*我是 ψ*这一思想的认知意义。其理由是，总是存在着这种可能：人们也许没有意识到我是 x 所以 Φx，所以即使人们会根据*我正在 ψ*(例如，*我正在把糖全都洒在超级市场的地板上*)的判断立即行动，也未必能根据*x 所以 Φx 是 ψ*的判断行动。

皮科克对比了一个人当他或她只是由于读过房地产商的宣传材料而知道墙的长度时所知道的东西，与一个人当他或她只是由于看过这堵墙而知道墙的长度时所知道的东西。弗雷格关于内容的直觉差异判据① 可用来说明：虽然两个人都知道墙的长度，但谁也不知道对方所知道的东西。因此，除了我知道墙的长度只是因为读了宣传材料，而我妻子知道墙的长度只是因为看见了墙这个事实以外，可以假定我妻子和我的认知状态是等同的。但是这样一来，由于我们每人只是以各自所能利用的方式想着墙的长度，我也许对*该长度大于我们的钢琴的长度*这一思想持不可知的态度(例如因为我们不知道我们的钢琴是多少英尺长)，而我妻子则判断这一思想为真，因为她只要看一下，就能知道靠墙放得下我们的钢

① 弗雷格的直觉差异判据：在一个认知行动中所掌握的思想 x，与在另一个认知行动中所掌握的思想 y 有差异，当且仅当某个理性的人在某一时刻有可能以不相容的态度对待它们；即接受(拒绝)一个认知行动，而拒绝(接受)或以不可知的态度对待另一个认知行动。

琴。所以感知的示范性β内容有别于任何以描述方式说明的概念内容，因而不能通过任何像"那个人看到以英尺计的距离(a,b)=n"这样的说明，其中a和b是墙的端点，以合乎概念内容的方式典范地加以说明。

我们可以用类似于我处理"想着斯坦福大学校园在这儿附近"的方式，也就是概念的方式，通过各自任务域的概念，来处理像佩里和皮科克那样的例子。这实际上是要以一个描述的、概念的方式去表征这些指示性内容。[1] 但是，佩里和皮科克的论证表明，以这样的方式不可能确当地对待这些内容的**认知意义**。所以我们只有识别一个以认知意义为要素的内容观念，才能弄清存在着一种不能通过任务域概念作典范说明的内容。

迄今的论证表明，存在着一大类认知状态(所有含有指示性或示范性元素的状态[2])，它们具有一种内容(β内容)，对这种内容来说，唯一的典范概念说明就是在共享感知环境或共享记忆经验的条件下，使用简单的示范性或指示性说明。对科学心理学的构造理论目标来说，这样的说明显然是无用的，因为理论家要理解这种内容的性质，唯一能使用的方法或是共享这种内容的经验环境，或是凭借理论家在记忆经验中所

---

① 参阅本书第517页注①，我在那里说，任务域观念摆脱了任何指示性观念。任务域中没有观点，所以如果观点对指示性来说是必要的，那么以指示性为要素的内容观念就不能通过任务域概念来获得。

② 是否存在着既不是明确地也不是隐含地含有指示性或示范性元素的表述状态呢？也许"上帝是善的"能算作这种表述状态，因为它是"上帝是独一无二的"这一本质的一部分。(差不多每一个确定描述都含有一个隐含的指示性参照，例如对我们的地球。)但是，"善"是指"对我们善"，"从我们的观点看是善，而不是比如说从魔鬼的观点看是善"吗？

能得到的类似的经验环境。[①](这里,科学心理学是以解决认知体现问题为目的的心理学,因而也是以从包含一些描述层次的非内容出发构造任何解释上不可或缺的内容观念为目的的心理学。[②])然而这一类别的内容对心理学来说是特别重要的,至少是由于它与行动的直接联结,以及它在学习中的决定性作用。因此,理论心理学家难道不能获得对我们在世界上行动和从世界中学习的能力来说具有基础作用的那些内容吗?

唯一条件是心理学家假定他或她必须以概念内容来工作。问题的出现是因为在示范性、指示性或观察性内容里面,不存在什么概念构造,可用来对那种会适合于科学心理学目的的内容作出一个典范的概念说明。但这并不排除在这种内容中存在着任何**非概念**构造。如果我们能够了解这一观念,那么这里就存在一个论据,可以说明心理活动大多只能通过非概念内容来获得,因而也就应该只根据非概念内容建立模

---

① 理论家可以在不运用所谈的表现方式的情况下指称所谈的表现模式,但这无济于事。如果心理学家只限于各种概念的说明,并且把需要构造任何心理学上不可或缺的内容观念当作解释性任务,那么所谈的东西就是,科学心理学家对这些内容的性质所能给出或使用的那种解释。麻烦在于如果这种说明是典范的,理论家理解所谈内容的性质的能力不可避免地依赖于他或她已经具有类似经验。这样,对于那些既是典范的又是理论上恰当的内容所作的概念说明就会失败,因为只有两种方法可以从概念上说明这些内容:一是通过任务域概念,一是使用指示性或示范性术语,即对术语使用的理解,或是依赖于共享经验环境,或是依赖于已具有的类似经验。佩里和皮科克的论证表明,第一种说明方法对β内容不可能是典范的,而不可言喻的对具有某些种类的经验的依赖则表明,第二种说明方法在理论上不可能是恰当的。屈森斯(Cussins 1990)文中详细阐明了其中原因。(感谢C·皮科克向我指出这种担心。)

② 我把过分严格的解释性要求施加于内容理论这种做法可能会遭到反对。我在另一文(Cussins 1990)中考虑了这种反对意见。

型。

基底域观念。于是我们放弃这种要求：每个内容必须以构成概念内容的方式即通过任务域的概念给出它的理论说明。还可能存在何种别的理论上恰当的说明方法呢？下面我介绍一种典范的非概念说明。虽然我认为这是我们能够据以解决认知体现问题的仅有的一种说明，但它未必就是仅有的一种。[①]

看一看一个生活在加里福尼亚州 SRI 研究院里的以“怪人”闻名的自主而可移动的机器人的运作，将是有益的。[②]“怪人”在 SRI 的走廊里行走，它的任务是沿着走廊来回走动，避免碰到墙壁，并转弯进入特设的出入口。

为了能够在一套任务域内灵活地行动，一个系统必须能够运用特征表述[③]，这些特征是它偶然发现自己在里面的那个区域所特有的。例如，如果在“怪人”所处的环境里，走廊的宽度是变动的，那么“怪人”就需要对走廊宽度作出不同响应。给出那种“怪人”系统，就意味着“怪人”必须表述这一变量。系统不需要表述的只是在系统的整个历程中没有变化的方

---

① 例如，如果德雷斯克的信息观念(Dretske 1981)要成为内容观念，那么它将是一个非概念内容观念，因为人们为了要处于信息携带状态，并不需要拥有信息的概念。(显然如此，因为即使是树——其实任何东西——也携带信息。)当德雷斯克试图证明把信息观念看作内容观念是正确的时候，麻烦就来了。皮科克(Peacocke 1989)发展了一个不同的非概念内容观念。

② 见赖费尔(Reifel 1987)。

③ 我不打算未经讨论就假定有了何种表述系统就足以拥有概念，所以在讨论像“怪人”那样的系统时，我用了一般的表述观念，该观念中立地看待是否它的意义(例如它的语义)只是外在属性，或者是否它的意义(像内容意义)是内在地可获得的这些问题。一个物理系统具有一些内在地可获得意义的状态的条件，我将在 § 7 中考察。

面。所以系统的表述能力越大,系统的潜在灵活性就越大。因此,我们是不是应该假定,在认知上是理想的系统,会以计算方式——以传统 AI 方式——表述所有存在的事实呢?是不是应该假定,虽然这个理想系统没有得到什么,但是人们越接近它,认知能力就将越强呢?

作出这种假定,[①] 就是忽视两种事实之间的重要区别。我要证明的是,这两种事实中,只有一种事实的计算表述是理想的人工智能系统所要求的。设想"怪人"有时运送皮萨饼经过 SRI。可能只允许一定重量的皮萨饼通过巨大的安全系统,因而"怪人"将根据这一假定来建造:如果某物被识别为皮萨饼,那么活动的手臂就必须施加一定的力把它举起来。这样做的结果将"暴露"出"怪人"在它每一次要举起皮萨饼时必须计算出需要多大的力来举起这张饼这方面的表述能力[②]。这种联结可以直接建在硬件中。然而,住在休里特-帕卡德公司附近的人们,被"怪人"日增的名声所吸引,也许想试一试让它为他们运送皮萨饼。他们可能会大大失望,因为对"怪人"说来遗憾的是,休里特-帕卡德公司实验室的安全系统允许所有各种重量的皮萨饼通过。人们发现"怪人"正在以不大可能引起 DARPA[③] 注意的方式把皮萨饼扔得到处都是。

的确,DARPA 可以据理争辩说,这是"怪人"的认知缺陷。我们以无止境的方式对待智能:某某可能擅长于下象棋,但是

---

① 正如某些 AI 理论家所做的那样,例如见莱纳特和费根鲍姆(Lenat and Feigenbaum 1987)。

② 见巴怀斯(Barwise 1987)。

③ 为"高级研究项目"服务的美国国防部基金署。

如果他不能学会玩牌戏“go”,我们就认为他缺少这方面的智能。对“怪人”来说,皮萨饼重量的表述是可接受的认知所必需的,更不用说理想的认知了。

但是我们不应该因此得出这种结论:“怪人”要成为真正有智能的,必须表示所有存在的事实。例如,“怪人”如果能表示出它底座上的声纳传感器之间的距离,人们反倒会感到惊奇。这不仅是因为这个距离在“怪人”的整个生涯中是一个定值,更重要的是因为“怪人”本身的结构并不是“怪人”的任务域的一部分。“怪人”从不需要操纵它的声纳传感器之间的距离;这个距离不是据以评价“怪人”性能的那种东西。相反它是“怪人”的能力基底的一部分,凭借于此,“怪人”才有了那些它实际具有的在走廊内移动的行为能力。任务域(t 域)与系统的能力基底域(s 域)之间的这种区别,对理解一个灵活的系统所必须表达的东西来说,是必不可少的。为了能够在一系列任务域内灵活地运作,一个系统必须能够表述一个任务域的那些在这一系列任务域内变动或可能变动的特征。但是只要基底域像通常那样,是在这一系列的范围之外,一个灵活的系统就不必去表述它的基底域的各个方面。[①]

我的视觉能力可能是极好的、不受限制的:我能在照明条件、离开我的距离等因素的广大范围内用视力辨别任何一种物体。但是任何人都不会指出:我不知道我的视觉信息处理系统所运用的算法,是我的视觉能力的一个缺陷。就我个人层次的视觉能力而论,我的下意识的信息处理能力是基底域

---

① 见屈森斯(Cussins 1987 年 5 月)。

的一部分。[1] 已知一个特殊情形中任务域与基底域的划分，任务域中的性能——甚至完全概念的性能——不要求拥有任何基底域概念。

用基底域概念说明 β 内容。正如我们看到的那样，任务域观念提供了 α 内容与概念特性的定义之间的联系。基底域观念能够提供 β 内容与非概念特性的定义之间的相应联系吗？一个智能执行者不必具有自己的基底域的概念，所以如果 β 内容能参照基底域的对象和特性而被典范地说明，我们就已经促成了一种通过系统或有机体不必具有的概念来说明的内容。

我们看看下面一段从埃文斯著作中引用的文字（1982：ch.6）：

> 一个受试者听到声音是来自空间中某某位置，这件事里面包含的是什么呢？……当我们听到声音是来自某一方向时，我们不必想或计算把我们的头（比如说）转向哪一边才能找到声源。如果我们确实必须这样做的话，那么就应该有可能两个人听到来自同一个方向的声音，然而在对声音作出反应时有十分不同地行事的倾向，因为他们的计算中存在着差异。既然这看来是讲不通的，我们必须说，具有空间上有意义的知觉信息，至少部分地

① 显然，何种能力是基底域的一部分，是与特定任务域有关的。视觉信息加工算法方面的知识不是许多人的任务域的一部分，但它是 D·玛尔的任务域的一部分。因此某个状态拥有何种内容，与适合于它的评价种类有关：即与特殊情形中划分任务域和基底域的特殊方式有关。对于单个状态，也许同时存在着不止一个任务域。

在于有做不同事情的倾向。

当埃文斯问"一个受试者听到声音是来自空中某某位置，这件事里面包含的是什么呢?"的时候，他是在问在经验中将世界的这个方面呈现给受试者的那个内容的性质。显然，这内容是指示性的或示范性的，因为要是用言语来表达这内容，我们就会说，知觉所呈现的声音是来自"那个位置"，或"从那边来"。以佩里和皮科克的例子为基础得出的结论是适用的：如果我们希望在理论上公正地对待内容的认知意义，特别是它与行动的直接联结，那么就无法将这一内容作为概念内容典范地加以说明。埃文斯补充的是，第一，关于为什么不能以概念方式获得这种内容的进一步的理由(任何一个概念内容都不能和某些β内容所要求的同样直接地与行动发生必然联系)；第二，关于获得内容的认知意义的方式以在世界中运动的方式为参照的建议，如受试者伸手去抓对象的能力，指出其位置的能力，或向声源走去的能力，都是由知觉经验得到的能力。在我们现在已经得到的论证的位置上，重要的是这第二个思想，因为对埃文斯的内容来说，在世界中运动的方式是基底域的一部分。

根据我们通常对意识的看法，这里的思想可能显得相当奇特，这一思想是：某些内容是在世界中找到个人的运动方式(比如说跟踪目标)的手段，即使主体不能在概念上获得这种方式，而且事实上主体也许不能用言语表达这个运动方式是什么，[①]却可以在他或她的经验中获得。我的关于声音来自

① 而且这个主体也许还是不能运动的。我有可能在我的经验中获得一种运动方式，即使我不能根据它来行动。(这种内容仍将根据它与知觉和行动的构成性联结得到典范表征。)

何处的知识在于比如说我如何确定这个地点的知识：通过我在经验中直接——不依赖于任何概念——获得的东西，就可以充分获得这一知识。即使我不能怀着任何有关当前所说的运动方式的思想，我也可以具有这个知识；我不需要任何有关我在环境中找到运动方式的能力的概念，就可以具有以为我提供一个运动方式为内容的经验。

考察最极端的非概念内容的情形之一也许是不无益处的（我在 § 8 中还将回到这一问题上）：疼痛经验的情形。我们已经了解到，在哲学传统中，全然不把疼痛经验看成带有任何内容的经验；我们被告知，它的功能不是表述世界。但是其理由不是因为疼痛经验从现象上看与关于物体颜色或形状的经验不很相像，而是因为我们不把世界看成拥有各种不同的疼痛特性。我们根据褐色的视觉经验说我的桌边涂成了褐色，但是我们并不根据剧烈疼痛的触觉经验说我的桌边具有剧烈的疼痛。我将在 § 8 中说明这种情况何以如此的原因，但是当前的要点是把经验看成一个经验种类的系列，从使我们往往倾向于把经验到的特性归属于世界的疼痛经验，经由使我们确实把经验到的特性归属于世界（但是我们因此而遇上一些麻烦—— § 6）的颜色经验，直到形状或运动经验。疼痛经验的客观性[①] 的确比形状经验差得多。这将显现在疼痛经验所能具有的内容种类与形状经验所能具有的内容种类的对比上。疼痛经验从不具有概念内容，但是不能因此认为它全然没有内容。疼痛经验将世

---

① 关于“客观”的意义我将在 § 7 和 § 8 中作出明确说明。大致说来，疼痛经验客观性较差的原因是它的视角无关性较差。

界呈现为是疼痛的；疼痛可以在疼痛经验中被人所获得。但这并不是说，我们是为了这种情况的存在而需要疼痛概念的；我们只是不得不处于疼痛中，或记得处于疼痛中。类似地看，即使经验主体没有任何运动方式的概念，经验也能够表现世界中的运动方式。

我们的动觉意识提供了另一个例子。根据动觉经验，主体知道他的身体是如何放置的；他的双手的相互关系如何，以及例如与他的头的相互关系如何。但是这个人不需要任何有关这一空间位置的概念，就能具有这种知识。相反，该知识在于对例如使双手靠得更紧或让双手紧贴躯干这样的运动的经验敏感性。人所具有的重新调整自己身体的能力，直接表现在动觉经验中，而无需拥有任何关于身体位置的概念。

我们回到听觉经验的例子，埃文斯的看法是，听觉感知的空间内容必须根据一套不是以概念作中介的能力来说明，才能形成判断和在有机体周围的自我中心空间[①] 中运动。这是因为内容是主体能从经验获得支配运动的能力。知觉的经验内容是根据有机体所拥有的某些基本技能来说明的，这些技能如："在视觉范围内跟踪一个目标的能力，或在复杂和发展变化的声音模式中追随一种乐器的能力"。这些技能都属于主体的基底域。所以如果埃文斯是正确的，这一类内容就是参照作为基底域的一部分的那些能力而典范地说明的，因此也就是通过主体在掌握这类内容中任一项时所不必具有的概念而典范地说明的。所以(在概念上是原子性的)指示性的、示范性的和观察性的经验内容的结构，就是它们的非概念

---

① 对这个术语的某些阐释，见 §9 中地图制造者的讨论。

内容的结构。β内容是非概念内容。

人们常把这种看法误解为行为主义理论，所以我再一次强调，首先这一论断并不是关于有机体的下意识[①]知觉状态的表征。其目的是获知个人的知觉经验是如何把世界呈现为存在(即真正的个人层次内容观念)的。非概念内容的观念是一个最终必须根据在经验中可获得的东西来解释的观念。如果内容被典范地表征为某一确指种类的复杂配置，那么该论断则是，这种配置是个人能在他或她的经验中直接获得的，并且经验内容就在于这种可获得性。但是对行为主义者来说，经验观念不可能具有任何解释作用。[②]

于是总括起来说，我已经认清了通过认知意义而不是通过真值条件对内容的制约；我已经根据弗雷格的精神提出，我们需要引入一种与这制约相应的内容；我已经证明了这种内容不能以任何合乎概念内容的方式典范地说明，所以它不是一种概念内容的类型；我们已经看到我们需要运用这种内容，以便充分公正地对待任何带有指示性或示范性元素的认知心理状态(我们认知活动的大部分)；关于如何以典范方式获取这种内容的一个可取的建议是利用基底域的概念；并且因为认知生物不需要具有基底域概念，所以我已经证明这种内容满足非概念特性的定义，因而是一种非概念内容。[③] 我要把

---

① 见丹尼特(Dennett 1969:93—4;Dennett 1978:101—2,153—4,219)。

② 其实不仅对行为主义者是这样。认知革命也许已经恢复了这种表述观念的地位，但它尚未恢复经验(我的内容观念)的地位。我希望这篇文章多少能推动我们朝这方面发展。

③ 应注意的是，也许有许多种β内容不是通过基底域概念，或者更狭义地说，不是通过在环境中找到个人运动方式的途径的概念，作典范说明的；要是这样的话，它们就将是并非概念内容种类的内容种类。它们的状况取决于打算以怎样的方式典范地获得它们。

所引入的这种非概念内容称为"构造理论内容(CTC)",因为我将继续证明,这种内容怎样才能形成构造概念能力的基础。

非概念内容的现象有多么普遍?

某些概念状态的内容仅有这些状态的非概念内容的结构,所以只能通过它们的非概念结构作心理分析。内容分析有两个层次,概念的和非概念的,我们某部分认知活动的心理学解释,只能通过它的非概念结构给出,这一点已得到证明。我们不禁要提出疑问,这现象有多么普遍?甚至对存在着概念结构的那些认知领域而言,正确的科学心理分析层次也仍然是通过它的非概念构造作出的,这可能吗?**认知的心理结构是非概念结构吗**?我相信认为它是的这一假设是 LOT 的联结论替换方案的基础。但是这一点超出了我们现有的认识。①

考察另外几个例子是有益的。埃文斯引用 C·泰勒的论述如下:

> 我们的知觉场有一个定向结构,一个前景和一个背景,一个上和下……这个定向结构标志着我们的场基本上是一个行为体现者的场。这个场的视角并不是集中在我身体的位置上——这本身没有证明我实质上就是行为

① 在我的另一文(Cussins 1990)中,我将这一论点从指示意义和示范意义的情形扩展到所有意义中去。

者。但是以该场的上下方向性为例，它是以什么为基础的呢？上和下不是仅仅与我的身体相联系的——上，不只是我头所在的地方，而下，也不只是我脚所在的地方。因为我可能是躺着的，或弯下腰去，或头朝下，在所有这些情形中，我的场中的“上”都不是我的头的方向。上和下也不是由这场中某些范式对象来定义的，像大地或天空：例如地面可能倾斜……事实上，上和下是与人在这个场中如何运动和行动相联系的。

这里，泰勒是在问：我们的概念*上*的意义是什么。他设想了三个答案，其中两个是：上是我头的所在地，上是天空的所在地。但是我们的上的观念的意义不可能在于我们对我们的头的方向或天空的方向的掌握，因为例如我们躺下时，我们能够完全正确地运用概念*上*，等等。于是泰勒提出第三个答案，“上和下是与人在这个场里怎样运动和行动相联系的”。这个答案由于与泰勒提出的前两个答案有很大不同，令人耳目一新。在前两个情形中，提供的是传统概念分析，是像我们可以定义“单身汉”来指未婚成年男性那样的定义。在宜于给出传统概念分析的场合，一个人对定义中左边的理解，必然在于展示在右边的概念结构所具有的认知可获得性。但是泰勒的第三个答案不是一个定义；它只说明，我们对概念*上*的拥有，必须通过我们所拥有的某些基本的、非概念的能力，例如我们以协调的方式运动和行动的能力来分析。这些基本能力可以利用技术概念(例如万有引力构成我们场时所用方式的概念)来表征，这些概念是有机体为了拥有这些基本能力所不必拥有的。泰勒已经发现了通过概念的非概念内

容分析概念的方法。

我们再来看一看那些(大多数)情形中的识别能力:(对任何具有特性*Φ*和*ψ*的概念来说)识别并不依赖于把对象识别为*x，所以Φx和ψx*。例如,我把我妻子的面庞识别为卡丽丝的面庞的能力,不是识别唯一一个具有某些概念特性(例如高鼻梁鹰钩鼻,眼距为n英寸……)的面庞的能力(甚至下意识的能力)。当我的思维中出现一个形式如*这是卡丽丝*的知觉示范性思想时,我的认知状态不能正确地重构为包含如下推理:那是人物Φ,人物Φ是卡丽丝,所以那是卡丽丝。事实上,当某人不在场时,我回忆该人(甚至是我十分熟悉的人)的特征的能力是极其有限的,但是这决没有降低我把一个特定人物保留在记忆中的能力。(在极端情形中,我也许回忆不起我妻子的单个的知觉特征,但是在形成对于她的独特认识方面,我的能力是无与匹敌的。)所以下述情况是不可能的:我按所要求的形成对某人的独特认识的方式将该人保留在记忆中的能力,在于我存储了某一套恰巧在世界中唯一地得到满足的概念特征。①

由此得到的启示是,我们的识别理论中虽然会存在心理特征,但是这些心理特征不同于这种特征:对该特征的分析取决于**语义**说明,即该特征与恰当的预先记录的任务域

---

① 记住*概念的*不等于*意识的*。当然,识别卡丽丝,并不依赖于与有意识地存储的特征相匹配;这不是我现在要表明的观点。在断言个体识别的心理结构是它的非概念结构,而不是它的概念结构时,我断言是个体识别的计算模型必须适合于变换具有非概念内容而不是概念内容的表述方式。当然,表述方式的这一计算变换的大部分通常是全然无意识的。

的某个客观元素之间的语义关系。我们的识别能力大大超过我们的回忆能力,同时也不能通过回忆能力来分析。[1]一个可供选择的具有启示性的观点是,我们回忆世界客观特征的能力,依赖于我们识别能力的非概念内容的结构;而这内容是通过基本的时空追踪和为在周围环境中找到我们的运动方式所必需的分辨技能来说明的。

非概念内容和表述载体

掌握了概念内容与非概念内容的区别之后,我们可以回到一般性论点上来:一个计算模型所运用的表述理论的种类,决定着适合于这一模型的内容的种类,后者又决定着这一模型所能提供的心理学解释的种类。在下一节中,我将考察要求概念内容的那种心理学解释,而在§6中,则将考察可通过非概念内容获得的那种心理学解释。但是首先在心理计算的层级体系中迈出一步,弄清楚两种内容与两种表述理论之间的联结,将是有益的。

我在§3中解释过:LOT回答认知体现问题的能力,取决于它使用了S/S表述理论。现在我们有可能在§3中认识的那些构成S/S理论的特征之外再增加一个特征。我已经指出,语义层次与句法层次在解释上是互相独立的,就是说人们不必了解语义学理论,就可以理解句法理论在说什么,反之亦然。句法必须顾及语义制约,但是在句法上定义的证明论或过程结果的运算是形式运算,因为它们与语义

---

① 见埃文斯(Evans 1982:ch.8)。

特征无关。到目前为止，我在谈到语义学时，只不过是说我们常常想使语义学理论成为一个可有限公理化的递归的语义特性理论。现在我们可以看出，语义学理论是关于句法条目与概念内容之间的关系理论。[①] 一个语义学理论之所以具有它所具有的那种形式，是因为作为基础的公理说明了任务域的对象、特性或情境的参照或指称的关系。S/S 理论是供概念内容使用的。

用于非概念内容的表述理论会取什么形式呢？首先要注意的是，载体与内容之间的关系不是由语义学理论给出的，因为这些内容是非概念的。载体携带的内容不能由载体与任务域元素之间的参照关系或满足（即语义）关系给出。其次要注意的是，CTC 的表述载体与 CTC 本身之间的关系不能是语义关系，因为 CTC 层次在解释上不独立于有机体能在它的周围环境内找到它的运动方式所凭借的基底域能力的层次。这些基底域能力是携带 CTC 的载体，[②] 但是我们看到，如果没有通过正是这些能力说明的这样的内容，我们就不能理解经验是怎样以非概念方式呈现世界的。非概念内容的"句法"和"语义"在解释上不是独立的，所以

---

① 严格地说，一个典型的 S/S 系统是不具有概念内容的，因为严格地说，它并不是一个概念运用系统。概念内容观念的这种用法是一种派生用法，它必须最终通过这观念是就一个经验和思想主体而言的范式用法作出解释（见本书第 512 页注①）。尽管如此，内容观念的派生用法是可能有很大用处的。

② CTC 载体理论有可能取许多不同的形式：控制论说明，信息加工说明，生态学说明（Gibson 1986），或张量理论说明（Pellionisz and Llinas 1979，1980，1982，1985），以及斯莫伦斯基解释（Smolensky 1988 b）。但是不应把内容表述载体理论与内容理论相混淆。我们不应说控制论或信息论内容，只应说内容的控制论或信息论载体。我在 § 9 中作出这样的区分：非概念内容的表述载体是能力，非概念内容的联结论计算载体是分布在加工单元上的激活模式。

严格地说，它们不是句法和语义。非概念内容的观念表明，怎样能存在一个彻底替换S/S表述理论的方案。这一观点的价值在于，如果我们要考察福多尔和佩利舒（Fodor and Pylyshyn 1988）对联结论的批评，即认为联结论的因果关系缺乏句法上的系统性，我们能够赞成福多尔和佩利舒的观点，但是我们要能证明这个事实如何产生对联结论有利的结果。

5. 概念论理论的形成和心理学解释

在计算上使用S/S理论势必要求使用概念内容，因而势必要求形成概念论理论。形成概念论理论是利用概念特性在**心理学**解释（见图15－2）层次上形成认知理论。如§4中定义的那样，将概念特性应用于形成有机体认知功能的理论，要求假定有机体至少拥有一套基本概念，据此得以建立由分析得出的概念层级体系[1]。这套预先假定的基本概念将被认为或是固有的，或是由非心理（例如神经生理的）过程获得的，或同是两者。

逻辑已为我们提供了一种理解存在于结构成分概念之间以及复杂思想之间的非时间性的逻辑关系的方式。认知科学中在知识表述方面的研究[2]已经产生了许多形式体系，将知

① “由分析得出的概念层级体系”只是意味着由于概念之间的单向逻辑依赖性关系而建立的概念顺序。因此概念＊眼镜＊在逻辑上或分析上依赖于概念＊眼睛＊，但反之不然。概念＊单身汉＊依赖于概念＊成年人＊，但反之不然。

② 例如见布拉赫曼和莱韦斯克（Brachman and Levesque 1985）。

识表述扩展到获得那种构成真正的理论推理和实践推理的动态时间关系。这样,一切心理活动都被看作是对基本概念(中央推理过程和学习过程)之间关系的各种不同的处理,或是有机体的中央概念处理系统与感觉及效应器系统之间的联结(一种通过周围神经系统模件、知觉模件和行动模件实现的联结;参阅 Fodor 1983)的建立。

通过预先假定有机体拥有一套基本的概念内容,理论家就预先假定了有机体可获得一个预先记录的世界(任务域),这世界由在语义学中被当做基本概念及其复合体的所指对象的一组对象、特性和情境组成。这一基本概念系统与世界之间的联结被认为是由周围神经系统模件实现的,这些模件除了某种参数设置之外,基本上是固有的。因此根据概念论理论,认知起点已经成为这样一个点,在这一点上客观(然而必须简单的)世界是可被认知主体利用的:预先假定客观性是为了使认知理论能够开始形成。如果存在着一个关于自然有机体怎样能获得客观记录这世界的记录能力(比如说,不同于以疼痛经验为基础的能力[①])的问题,它不是被当做一个心理学问题,而是被当做一个神经生理学问题,或自然选择理论问题。

与此不同的是,因为拥有 CTC 并不取决于拥有那些 CTC 借以得到典范说明的概念,所以非概念论理论的形成不需要预先假定有机体拥有基本概念,因而不需要预先假定有机体可获得预先记录的世界。相反,形成非概念论理论(正如我们将看到的)就是形成关于那些使有机体可获得

① 见 §4 和下一个脚注。

客观的[①]、被记录的世界的过程的理论。所以关于如何获取基本概念的问题并没有被抛在一边让别的学科去研究，而是被直接作为心理学的中心问题来对待。正如我将在§6中解释的，在形成非概念论理论的过程中，把认知当做**客观性的显现**，而不是对客观性现成记录的推理处理。这样做的一个结果是，知觉和行动系统不是以建立体表刺激与中央推理系统之间的联结为功能的周围神经系统模件，而是认知的核心。高层次推理只不过形成了围绕这一核心而建立的周围结构。

形成概念论理论的另一后果是，在基本概念的解释层次之下，没有任何心理学的东西；只有实现方式[②] 理论。这是一个关于概念活动如何在非概念过程例如神经生理学过程中得到例示说明的理论。认知心理学中的概念理论可能在计算过程结果理论中实现，正如在逻辑中，语义(必要关系)是在句法(证明论)中实现的。问题是，对形成概念论理论来说，位于基本概念之下的东西是**在解释上独立于**概念层次的，就是说人们能够充分理解在概念层次上发生着什么，而无需理解关于这一层次之下正在发生着的任何东西。[③]

---

① 读者可能已注意到，我正越来越多地使用“客观的”这个词。一个相当粗略的想法是，如果某个事情独立于主体对它的掌握，那么它就是客观的。疼痛特性在这个意义上不是客观的，大学学位(?)也不是，而呈三角状是客观的。我将在§7中就这个观念谈谈比较精确的想法。

② 见本书第501页注④。

③ 哲学家们注意：概念论者所说的层次之间的解释独立性是与并发现象相容的。它是否与构造制约相容，则是一个困难得多的问题。我最终认为是不相容的，但是证明这一点取决于证明概念论在构造概念方面的所有尝试如LOT和机能主义都失败了。福多尔反对概念和非概念层次在解释上的完全独立性，因为例如他认为概念层次上的隐晦性现象必须参照认知表述的句法形式来解释(“命题态度”，见Fodor 1981)。

形成概念论理论的口号是，在概念之下的东西是在解释基石之下的。

由于局限在对认知的概念结构的关注上，概念论者不得不把所有的认知过程作为示范性的或非示范性的推理过程来建立模型（甚至学习也被作为形成前提和证明或否证前提来建立模型）。事实上，概念结构恰好是那种为了建立某个分析所需全部逻辑推理模型而所需的结构。所以心理学被认为是在解释上依赖于逻辑的。一个概念是什么，是由它在推理网络中的位置和它与世界在心理学方面的外在联结来确定的。与之不同的是，形成非概念论理论利用了原子概念**内部**的非概念结构。一旦单个概念的形状经过正确建模，推理组合的种种可能就会随之成为（并可解释为）这些元素的嵌套形式。〔在随后的论证中，我将集中论证内容的系统性（§7）。如果我们能证明有可能在非概念内容层次上获得系统性，那么就已经证明了有可能在非概念内容层次上获得推理联结。[1] 得到正确的元素形状使正确的组合嵌套形式随之而生，这会是何种情况呢，这一点将在§7中考察。怎样才能实现这种情况，将在§8中考察。〕

这种看法是不正确的：建立认知的组合结构的模型势必要求建立带有句法和语义表述系统的认知模型（Fodor and Pylyshyn 1988）。因为如果非概念论的思想是正确的，就有可能通过在概念内部建立非句法/非语义表述结构的模型而获

---

① 论断是：推理只是内容系统性的一般现象的特殊情形。推理是思想内容的系统性。

得组合结构。不应取这种看法:认为概念的性质在解释上依赖于概念之间推理联结的性质,而应是:推理联结的性质被解释为构成性概念的非概念性质的后果。

作为总结,下列要点对形成概念论理论来说是内在的:

1. 预先假定基本概念;

2. 概念与世界的联结是以固有性为主的"周围神经系统"联结;

3. 预先假定客观性:认知起始点假定了有机体可得到的一个任务域;

4. 获得把世界记录为客观世界的能力的问题不是一个心理学问题,而是一个神经生理学问题或自然选择理论问题;

5. 确定心理学解释基石的位置:形成心理学理论的最低层次是处理基本概念。概念特性仅仅以非概念特性来实现。

6. 所谓表述结构,首先是概念之间推理联结的结构。是概念的性质通过它的推理联结的性质来解释,而不是相反。解释的中心问题是推理,而不是学习。[①]

这些要点中的每一点都是把心理学解释专门局限于概念内容而产生的后果,所以也是在计算上和心理上应用 S/S 表述理论的后果,因而是 LOT 的一个结果。假定我们想要用心理学来解释自然有机体变得能获得基本概念,假定我们想用心理学来解释有机体怎样拥有思想元素(假定我们无论出于

① (1)和(4)得到福多尔(Fodor 1976,1980,1981a)的认可,(2)得到福多尔(Fodor 1983)的认可,(3)像其他要点一样,隐含在大部分 AI 的实践中,(5)和(6)为福多尔和佩利舒(Fodor and Pylyshyn 1988)所利用。

什么理由，想要没有(1)—(6)这些后果的心理学解释)，那么我们就需要一个 LOT 的替换方案。如果我们假定心理学解释必须以**某个**内容观念为基础，并且 LOT 的替换方案必须仍然是一个计算理论，那么我们所要求的理论的基础就是在心理上和计算上使用非概念内容。[①] 但是这样的心理学解释**可能**是什么样的呢？怎样能有一个取代概念论心理学解释的方案呢？

用非概念内容(CTC)的观念武装起来之后，我们怎样能公正地对待认知的概念特征呢？如果一个认知活动只能在非概念内容的层次上得到表征，它就只能包含对世界的十分原始的记录。然而即使是**原始**的世界记录，也只有在可能把它展示为深奥的或充分概念的世界记录的一个简单形式或结构成分时，才是一个原始的**世界记录**。我们怎样才能做到那一点呢？

6. 非概念论的心理学解释：客观性的显现

概念论理论的形成为它的语义学理论预先假定了一个记录现成的世界——一个任务域。因为 CTC 是通过基底域而不是任务域的特性来说明的，所以基于 CTC 的心理学理论

① 我们并不**因此而**需要一个以 CTC 的心理学应用和计算应用为基础的理论，因为可能存在别种非概念内容。但是我贯穿全文的宗旨是指出某种理论是如何可能的，而不是指出它是如何必要的(即使我相信它也是必要的！)。

的形成不必预先假定一个记录现成的、完全客观的、对主体来说是可得到的世界。所有预先假定的东西是基底域的基本的、非概念的有机体能力。非概念内容在与背景概念能力相分离的情况下(当然)将这世界呈现出来,但还不是客观地呈现为一个从任何视角都能得到的世界,同时人们对它的认识也可能出错。[①]

参照领域在解释上先于认知领域吗?

在经验内容**仅仅**是概念内容的情况下,客观世界的某个元素是**作为**(并且仅仅作为)客观世界的元素(因而作为任务域元素)给予主体的。由于这一原因,对原子概念内容的恰当说明,就是使用与世界中作为内容所指对象的客观元素具有原始语义关系的任何语言条目(名称或谓词)。(同时分子内容可以作为来自原子内容的逻辑结构体来说明。)这样,如果我把交通灯感知为交通灯(而不是,比如说,一个细长的有色标的糖果),那么对于我的经验的这一方面的内容的恰当说明,就是使用短语"交通灯"。这就是内容的**所指**说明。它选出交通灯作为任务域元素。[②]如果经验的内容仅仅是概念的,那么就可能采用经验内容的任一方面,并假定这个方面把世界的客观元素呈现为客观(即任务域)元素,然后把这个元素命名为内容的所指对

---

① 我将在§7中解释其原因。

② 这样,就将这一内容同细长色标糖果内容区分开来,因为在任务域中,细长色标糖果是不同于交通灯的客体(虽然它们有可能占据同一位置)。这是因为任务域是**在给定概念化之下**的世界中的一个区域。

象。

但这是不可能的。我们考察一下与一个可能理论背景相脱离的、特殊的颜色经验。[1] 假如我们试图参照性地说明这一经验的内容是**对于**特殊色度的经验，我们其实是以为我们的经验受到下面两个原理的支配[2]（以及其他原理的支配）：在主体对于表面 A 的颜色经验与他或她对于表面 B 的颜色经验无法分辨出不同，而主体的知觉功能正常地和正确地发挥作用的地方，表面 A 的颜色等同于表面 B 的颜色；在主体对于表面 A 的颜色经验与他或她对于表面 B 的颜色经验可分辨出不同，而主体的知觉功能正常地和正确地发挥作用的地方，表面 A 的颜色不同于表面 B 的颜色。但是因为众所周知，基本颜色经验是一个尺度，这一尺度上的不可分辨的差异是非传递性的，所以我们的尝试很快就会导致矛盾，因为我们会得到这一结果：A 的颜色等同于 B 的颜色，B 的颜色等同于 C 的颜色，但是 A 的颜色不同于 C 的颜色。因此，A 的颜色等同于又不等同于 B 的颜色。所以基本颜色经验不能参照性地用熟知的概念论方式来说明。根本不存在精确的色度，通

---

① 我所谓离开特定理论背景的颜色经验是指离开主体的某种知识来考察的颜色经验，这种知识的一个例子是人们可能根据颜色与所有对象的匹配条件来分辨颜色（Goodman 1951）。这种基本的颜色经验受到两个原理的支配，这两个原理是我根据皮科克（Peacocke 1986）著作中的观点编写的。（因为这些原理参照了色觉的因果条件，所以**基本**颜色经验不是**原始**颜色经验，而原始颜色经验是离开了主体认为经验是以某种方式从因果性中产生的这一知识来考察的经验。）高度的逻辑复杂性可能引起对颜色内容的连贯参照，但是这些颜色内容将不再是基本颜色经验的颜色内容，而是超复杂的集合论概念的内容。基本颜色内容本质上是观察性的：通常人们只要通过看，就可以知道它们是否适用于世界。

② 皮科克（Peacocke 1986）文中给出了这一论据，它利用了达米特（Dummett 1978："Wang 的悖论"）文中的论据。

过参照它们,我们就能够说明基本的、观察性的颜色经验的内容。①

这里我们必须仔细一点,因为(或多或少)存在着一种可以作为基本颜色经验内容的概念*红*——我们认为世界在客观上是红颜色的。所以我们能够说明颜色经验是与对于红色表面的经验一样的。然而对概念论者来说的问题是,我们是否能预先假定一个关于何谓表面是红色的解释,以便通过与被假定条目——红色色度——的语义参照关系,用熟知的概念论方式来表征基本颜色经验的内容。但是前面的论证表明,能起这种作用的红色色度是不存在的;如果存在红色色度,它们是不能成为任务域元素的。因此,解释表面**成为**红色是怎么一回事,不能先于解释**经验到**表面是红色的是怎么一回事。(因而,关于我们的颜色经验的内容理论不能**预先假定**一个具有客观颜色特性的世界;它不能预先假定自己的任务域。用概念*红*说明一个内容,是用概念的、而不是概念**论**的方式对内容作出说明。因而将由非概念论者对概念归属作出恰当的解释。)

为弄清楚这一论证的结论所作的尝试,揭示出形成概念论心理学理论所隐含的推论:概念论者通过预先假定心理上拥有基本概念,从而预先假定了一个关于何谓世界中存在着一个作为基本概念合适指称的客观所指对象的心理上独立的

---

① 以人为方式定义出无矛盾的颜色参照物,在这方面已作过许多尝试。其中有些也许是成功的,但是它们都包含某种对我们直觉实践的背离。人为的参照物将不是由我们的基本颜色内容的认知意义来确定的(因为感觉要确定参照)。正如本书第547页注①中提到的,我在这里谈的只是基本的、观察性的颜色经验内容。

说明。这也许是物理学要告诉我们的,但是如果是这样的话,心理学在这方面就不能起到任何解释作用。通过预先假定拥有基本概念,概念论者是在假定,从心理学的视角看来,关于什么是基本概念的本质所在(除了它们与其他概念的相互作用以外),我们能够说的仅仅是,它们的内容在于与世界的一个客观条目的语义关系。所以概念论者根本不可能从心理学角度说明,存在着基本概念的所指对象是怎么回事。为了从心理学上作出解释,我们假定出这一世界,并试图通过它去解释心灵的性质;也就是说,世界对于心灵,有一种解释上的优先权。

但是来自颜色经验的论据恰恰对这一解释优先权提出了怀疑。(不管怎么说,非常令人怀疑的是:我们真的设想物理学能够不顾及心理学上对像我们这样的有机体将某物识别为椅子或……是怎么回事的解释,告诉我们椅子是什么,英式足球赛是什么,大学学位是什么,或弄皱的衬衣是什么吗?)

(以贝克莱的方式)采取相反的优先权同样是似乎不合情理的。较为令人满意的方案是,假定对认知的解释和对世界的解释是互相依赖的。如果把这种想法与认为我们应该使用非概念内容的观念来提供我们概念能力的构造的想法结合起来,结果就会隐隐显出非概念论的心理学解释:思维能力的情况就是**客观性的**非概念**显现**的情况。对非概念论者说来,心理学解释的元素不取决于心灵/世界区别对于待解释词"主体"或有机体的适用性。因此这些非概念元被用来说明这种区别怎样变为适用的,从而用来说明待解释词有机体成为经验和思想的主体是怎么回事。

心灵/世界区别的显现

我来简单谈一谈这个标题的意思是什么。我们平时谈到认知时嵌入了一个心灵/世界的区别;我们把我们的认知状态表征为是关于一个独立于心灵的世界的,所以我们是通过两个独立存在的实体——心灵和世界——之间的外部关系来表征认知的。我指的意思不仅仅是:即使没有任何东西在感知世界,世界也继续存在,或者世界的真实情况比任何人知道的都多,或者存在着我们没有能力识别的真相,而且我还指这样的意思:在心灵与世界之间存在着一道鸿沟,使心灵对世界产生误解,对世界有所指(而不是仅仅像草履虫一样浸没于世界中),因而能产生关于世界的思想。

这件事的一个表现形式是,我们对两种谓词作出泾渭分明的区别,一种谓词能恰当地与对世界有所指的主体词相结合,一种谓词能恰当地与对心灵有所指的主体词相结合。因此,我们能说一个球是圆的,而不能说一个经验是圆的。我们能说一个记忆是诚实的,而不能说一个足球场是诚实的。在同一个谓词能用于两种语境的场合,我们坚持认为,严格地说,它在每个语境中的意义是不同的,或者在一个语境中用的是本义,在另一个语境中用的是比喻义。记忆和足球场都可以说成是笼罩在烟雾中,但是……

在这一点上有少数几个例外,其中之一是谓词"……是客观的"。经验和形状特性都可以是客观的。这一谓词怎样能具有一个特殊地位呢?假定心灵/世界区别是种系发育或个体发育的**成就**:草履虫没有表现出这种区别,但是我

们表现出来了,幼小的人类婴儿没有表现出来,但是成年人表现出来了。自然选择通过逐渐而连续的遗传变化,进化出具有心灵/世界区别能力的人;婴儿内部的学习过程通过逐渐而连续的神经生理学变化,发展出把独立的世界呈现给独立的心灵的认知能力。于是问题出现了,在心灵/世界区别得以应用之前,我们能够借助定义在非概念内容之上的过程理论来描述、解释和了解发展的前客观阶段吗?我们能够借助以计算方式处理非概念内容的理论,解释从这种未拥有任何概念的前客观阶段向运用概念的客观阶段的过渡吗?非概念论的心理学解释正是试图这样做的。因此,这是试图证明心灵/世界区别——客观性——怎样能从仅仅存在着一个未分化的心灵/世界连续体的前客观阶段中显现出来。心灵是**嵌入的**(Cussins 1987 年 5 月),而不是唯我论的或形式的(Fodor 1981 a),因为一个理论可以算做拥有内容的理论,当且仅当它也是关于何谓概念所指对象存在的理论。

当然,我们的描述和解释全都是从我们的概念视角出发的,所以我们能够从心灵的视角,或世界的视角描述未分化的心灵/世界连续体。暂且采用前一视角。于是客观性的显现就是从纯粹的经验到**对于**世界的经验的转变。① 再

① "我们能够想象一系列的判断:'现在暖和','现在嗡嗡响',这是主体在响应他的感觉状态变化时产生的,根本没有客观意义。但是我们可以想象由主体感觉状态中相同变化激起的一系列类似的判断,它们确实具有这样的意义:'现在是暖和的','现在有一个嗡嗡的声音'——这些都是对一个变化着的世界所作的评论。在意义的这种变化中包含着什么呢?"(Evans 1980)。斯特劳森(Strawson 1959:ch.2)、埃文斯(Evans 1980)和斯特劳森对埃文斯的回答(Strawson 1980)之间的交流,是对客观性的显现采用心理视角的经典讨论方式。

取后一视角。于是客观性的显现就是从原子在晶格空隙中飘移到熨烫皱了的衬衣和设计合作处理文件的办公环境的转变。

论断布景[①]：来自并发现象和上层结构的反对意见

人们也许会假定，学习——客观性的显现——只不过是一个爬上后就被踢开的梯子。换句话说，概念论者也许承认我们需要对种系发育和个体发育的发展作出非概念论解释，但却坚持，一旦发展起来，成年人的认知就适合以概念论的方式来对待。但这样将使我的论证前后倒置。我不曾认为，我们需要就学习作出非概念论说明，因而我们需要就成年人的认知作出非概念论说明。相反，我已认识到必须从心理上识别成年人的认知中的非概念内容，从而认清了这样引入的观念适合于对学习的解释。

学习观念可具有两种意义，这取决于是将观念嵌入概念论理论，还是嵌入非概念论理论。如果是前者，那么学习就只是一个将被踢开的梯子，因为通过爬上梯子所取得的某种东西，它的性质是由独立于学习理论的理论（推理理论）来解释的。但是如果是后者，那么对（成熟的）认知来说，学习就是必不可少的，因为所学东西的性质在解释上并不独立于学习机制。对非概念论者来说，学习是客观性的显现，而（没有学习

---

① 这里有一个小小的舞台布景，基本上是以比喻方式来展示某些论断和反论断的。

的)认知是客观性的保持。用来解释保持客观性的解释性观念,与学习理论中运用的观念是相同的。按概念论者赋予学习观念的那种意义作出假定,然后论证说因为C3假定学习对认知来说是必不可少的,所以C3必然是错误的,这样做是不合理的。

我的论断是,我们虽然是成年人,却仍飘浮在部分分化的、部分客观的心灵/世界连续体中,在这个连续体中,疼痛经验居于一端,稍前进一步是各种情感经验,然后是颜色经验,然后也许是形状经验和关于民主公正的经验。在我们进行认知活动的所有时刻,我们对概念的拥有或者是处于浮起状态(形状经验),或者只是半沉溺状态(颜色经验的悖论),或者是几乎完全沉没的状态(疼痛经验)。从本义和比喻义两个方面来看,概念的运用都是我们在环境中找到我们道路的能力的结果,即处于浮起状态。我们的非概念能力**支撑**着我们的概念能力;它们不仅是转变阶段的一部分而已。

概念论者的反对意见可能采取两种形式。第一,他可能假定在有机体的形式中存在着一个C3的反例,由于宇宙的巧合,这个有机体重现为一个构造完整的成年人,也许他的分子与以正常方式出现的人的分子有着等同的类型。会不会用C3来否认这样的生物拥有概念,这种否认会不会与并发现象不相容?[①] 但是在断言学习的中心地位时,C3并不认为一个重现的人不会是一个概念拥有者(如反对者所提出的)。只是认为,对我们同意他会拥有的概念能力作出**解释**,要借助学习理论。我的论断是关于解释方向的论断,而不是关于并发现

① 这是N·布洛克在伯克贝克讨论期间提出的反对意见。

象的本体论的论断。概念论者的论断也是如此,他有可能承认一个从未从事推理的概念拥有系统的理论上的合理性,同时也认为对概念的拥有作出解释要利用推理理论。所以非概念论者能断言关于何谓拥有概念的解释要通过学习理论给出,同时承认可能存在着从未从事于学习的概念拥有系统。不过在这两种情形中,只要概念得到**运用**,拥有概念的过程(前一情形中是推理,后一情形中是学习)也就得到执行。

概念论者可能提出的第二种形式的反对意见,必须涉及认知的上层结构,例如那些含有语言的认知能力。反对者虽然愿意承认语言能力是以非概念能力为**基础**的,但仍然坚持认为,一旦这种基础存在,它们就会自行增长,产生出适用LOT的认知的较高方面。[①] 为简单起见,在本文接近结束时,我较少用比喻方式谈及语言认知,但是按照论断布景的精神,我在此特别提出C3回答的形式。

我们必须留心记住载体与内容的区别。语言认知和在语言上受影响的认知,对于人类认知的科学心理学是至关重要的,但是我们的论证不应该是从语言载体在人类认知中的中心作用到LOT式的理论,LOT式理论假定,在传统上通过S/S理论与语言相联系的**概念内容**对于人类认知的科学心理学来说是基本的。的确,我曾论证过,语言载体常表达**非概念内容**。非概念论者尽管承认语言载体的重要作用,但却坚持认为:是这些载体的非概念内容,而不是它们的概念内容,在心理学中起着重要作用。非概念内容在构成认知基础方面并没

① 这里我想到A·克拉克和M·戴维斯提出的反对意见。并不是说他们就是概念论者!

有达到为构造认知上层结构提供建造模块的程度。重要的问题是，在科学心理学回答认知体现问题时，是否需要 S/S 理论。概念论者认为需要的论断，不仅是语言是表述载体在心理学上的基本形式的论断，而且是 S/S 理论是正确的语言理论的论断。恰当地识别语言在认知中的作用，会使这两个论断都成为不必要的。

单词确实常常表达概念，它们的这种作用，对我们的认知活动有着极其重要的意义。C3 并不否认这一点，首先，C3 的目的是要表明，这一现象的 S/S 理论怎样成为无用的：因为它不能用理论上恰当的方式获得指示性和示范性内容的认知意义，因而它被错放在产生恰当的学习、知觉、记忆和行动（它们本质上都是指示性的）理论的位置上。其次，毕竟更重要的是，有一种处理语言概念现象的方法，它不是依赖于 S/S 理论，但从自然主义的角度它是可以接受的。我们需要确认概念的重要性（§7），然后说明非概念内容的计算理论怎样能获得这种重要性（§8 和 §9）。[①]

虽然存在着颜色概念，但是由于认知意义的悖论和问题的诘难，基本的颜色认知不能用概念论的方式来解释。非概念论者假定，我们的认知活动全都与基本的颜色经验相像，只是稍好一点或稍差一点；通常我们可以用概念方式处理它，但决不能用概念论方式。概念内容是有价值的理想化，而不是心理学解释的基础。显然，与其说它是数字认知方面的理想化，不如说它是颜色认知方面的理想化，但是我们的认知的广

① 此外也有一些重要的认知现象是依赖于语言载体以及别的交流载体的。对这一点的认识完全在 C3 精神的范围内（§9）。

泛基础与颜色的情形并没有太大的不同：我们可以把认证从表面颜色经验的不可分辨的差异的非传递性推广到适用连锁悖论的任何模糊概念：*秃的*，*一堆*，许多形状概念(例如我有一个25边形的概念，但不是**作为**一个25边形)，字母"A"的概念(不能从几何学上或拓扑学上对它作出说明)，音位〔p〕的概念(不能从声学上对它作出说明)，可从伦理学上和美学上评价的概念，民主的概念，我的关于通心粉的概念……(如果有什么例外的话，则是数学或集合论的概念。)

概念内容在任务域(的一部分)经验中是可获得性的。CTC内容则在基底域能力经验中是可获得性的。我们看到，非概念内容不能借助任务域说明来解释，所以非概念内容不能用概念内容特有的方式来解释。**非概念论的心理学解释取决于相反的可能性**：概念内容**能够**用非概念内容特有的方式来解释。概念的运用——我们知道就是任务域经验中的可获得性——原来是基底域能力经验中的可获得性的一个特殊情形。对非概念论者来说，基底域能力的经验可获得性的观念，是对任务域的经验可获得性观念的概括。因此，科学心理学应该放弃这个一般性较差的观念，而完全通过CTC内容特有的理论机构建立认知模型。但是这样一来，我们就需要对基底域的经验可获得性构成任务域的经验可获得性的条件作出解释。

7. 客观性制约

如果非概念内容要在非概念论心理学解释中起到我所指出的那种作用——作为逐步构造客观性的基础，那么

我们就需要理解，为什么把某些非概念认知状态解释为向主体呈现客观世界是不恰当的，以及为什么某些别的更复杂的非概念状态又确实可以算作是提供了客观的记录能力。

概念论者几乎不需要谈及客观性①，因为心灵与世界之间的客观关系是他形成心理学理论的先决条件。但是非概念论者就承当不起这种奢华的做法；因为对他来说，心理学解释是对于从一种前客观状态转变为一种客观状态的解释。我们需要理解，在草履虫这样的动物与我们这样的动物之间，在婴儿与成人的认知之间，也许还有在正常的与精神错乱的认知之间，怎样能存在一种即使不是鲜明的但却是原则性的区别。在充分掌握这一区别的原理之后，我们就能在§8和§9中讲出，如何最好地建立能够把一个生物从处于这区别的一侧转换为处于另一侧的计算过程的模型。

关于客观性的某些直觉认识

我们关于世界的观念是那种不依赖一个有机体与世界的特异关系的观念，因为它是关于全部有机体所共有的东西的观念。如果这些特异关系是以信息方式说明的，那么我们关于世界的一个元素的观念，就是某种不依赖于任何与它的特殊信息关系的观念。如果这些特异关系是以经

① 我们的确很少听到心理学对客观性的讨论。发展心理学的某些领域，特别是皮亚杰心理学，在少数例外情况之列。

验方式说明的，那么我们关于世界的一个元素的观念就是关于某种不依赖任何对它的特殊主观经验的观念。各种不同的表述系统可能对世界取十分不同的视角，然而我们之所以能够谈论“视角”，仅仅因为存在着一个共同的焦点，某个可以用十分不同的方式来观望的共同的事物，一个用来表示同意和不同意的共同基础。客观性的理论是一个形而上学理论，所谈的是何谓在这种意义上存在一个世界。正是这一理论谈到了，何谓一个有机体与世界的特殊关系的意义超越了这些关系中的特殊事物，超越了特殊的能量构形，或感觉资料的特殊构形。如果关系状态中存在的一切只不过是具有这种能量构形而已，那么处在关系状态中这件事就不能“呈现”任何超越特殊关系本身的事物。所以它就不能呈现某种对全部有机体来说是共同的东西。因而，一个客观性理论就是在一个有机体与世界关系的特殊性和世界本身之间进行“认知”分离的理论；正是这一理论谈到了何谓存在着个体认知过程与共同的世界之间的区别。

探讨何谓客观性和何谓客观地记录世界的一个方法，是询问为什么我们会认为这种认知分离过程表征出人类与世界的关系，而不是草履虫或青蛙与世界的关系。以恒温器、伏特计、草履虫和青蛙为一方，以进行认知的人为另一方，什么是区分它们的判据呢？我们能给出以概念方式对常识世界作出的响应与换能器对某种特性作出的响应之间的在理论上有牢固基础的区别吗？[1]

---

[1] 见福多尔(Fodor 1986)。

客观性的连贯性试验：青蛙和自动螺丝刀的情形

我建议对这区别做一个试验，这试验是以心灵和世界的解释是互相依赖的这一见解为基础的。这试验本质上是这样的：把功能系统的能力看作是用非概念方式描述的，并尝试把这些能力解释成概念的。这一尝试将是成功的，当且仅当可以借助（假定存在的）概念能力呈现的（假定存在的）世界是一个连贯的世界。我们使运用概念的有机体与运用非概念的有机体的区别以关于世界的连贯性条件或我所说的客观性制约的形而上学理论为依据。①

我们来看一个简单的例子：一个自动螺丝刀配有某些简单的感觉装置，如我们粗略谈到的那样，它能探测螺丝是否存在，如果存在，那么它是拧紧的还是拧松的，如果我们把概念（因而也就是把客观地记录世界的能力）看作是这个螺丝刀的属性，这样是否正确呢？这一探测螺丝过程的说法必须用工具主义的方式来解释呢②（Dennett 1987），还是它本质上与我们看作是基于概念的、我们自己的知觉机制的那种实在论属性的说法是同类的呢？在应用这个试验时，我们首先要问的是，呈现于一个带有这些以非概念方式表征的能力的系统面前的世界，会是什么样子的呢？（它会满足对客观性的制约

① 这一方法像埃文斯的方法一样，受到斯特劳森（Strawson 1959：ch.2）的影响。

② 这是一种有用的但不真实的说法。说天气很糟糕也许是有用的，尽管严格地（即真实地）说，天气不是那种能够变糟的东西。人们常常认为，把信念看作是恒温器的属性，同样是工具主义的做法。丹尼特（Dennett 1978）认为，即使把信念看作是人的属性，也是工具主义的做法。

吗?)我们可以尝试把这些能力解释为概念的,并回答说这世界只能是一个含有螺丝和拧紧或拧松两种特性的世界。但是这样的世界不是连贯的,因为螺丝只能是这样的世界的一部分:在这个世界中有制造它们的工厂,有它们表现出的刚度、长度和重量的特性,有它们所在的位置,有拧它们的方向,……所以自动螺丝刀根本不能算作是拥有任何概念的。

这里表明的观点是:由于人工制品在功能上定位于人类世界之中,所以关于与螺丝刀相联结的螺丝概念的说法是外在的说法。我们是通过人类世界的元素,以工具主义的方式来表征螺丝刀的能力的,因为我们能愉快地为了设计和评价人工制品的目的而预先假定人类世界。我们能够说这个人工制品在"探测螺丝",仅仅因为我们能预先假定这一点。所以对螺丝刀能力的理解,根本无助于理解某些物理系统怎样才能以概念方式获得世界。

类似的观点适用于我们的对比如青蛙行为的描述,这里我们关于青蛙探测苍蝇的说法,仍然取决于概念的外在属性,我们接受这一属性是为了理解青蛙的进化"设计"。我们再一次预先假定人类世界,因为只有这样做,我们才能十分恰当地预测青蛙的行为,十分恰当地理解青蛙设计的成功,而不至于变得淹没在心理细节,或纯信息论说明中。对青蛙来说,没有什么内在的和非工具主义的概念属性,因为任何把青蛙的以非概念方式表征的能力解释为是以概念方式呈现世界的尝试,都会产生一个不连贯的世界。如果没有大小尺寸,任何东西都不可能[1]是

---

① 对哲学家来说,此处可能性和必要性的观念植根于描述性的形而上学(Strawson 1959)。

苍蝇，但是青蛙探测系统的成功，并非取决于将近处的苍蝇与远处的庞然大物分辨出来。如果不能处于静止状态，任何东西都不可能是苍蝇，但是青蛙探测系统的成功取决于苍蝇的运动。青蛙关于苍蝇的观念是一个关于某种始终运动的、没有尺寸的东西的观念。所以这不是一个关于苍蝇的观念。所以青蛙没有苍蝇概念。①

整体论制约和普遍性制约

这些例子引出了对于客观性的**整体论制约**：任何东西都不可能算作对象或特性的概念，除非它是一个复杂的、整体的概念网的一部分，因为任何东西都不可能是这样一个对象或特性，除非它是概念系统所指的这一复杂的、整体的对象网和特性网的一部分（反之亦然）。

整体论制约加强了埃文斯的普遍性制约：一个有机体不拥有关于一个对象的概念 * a *，除非它能对于它所拥有的（并且在与 * a * 的结合中不会出现语义异常的）那些特性的所有概念 * F *，* G *，……认为 * a 是 F *，* a 是 G *，等等。类似

① 反对者也许同意青蛙没有我们的苍蝇概念，但仍坚持认为它们具有青蛙的苍蝇概念；或者，自动螺丝刀没有我们的螺丝概念，但它确实具有"黑色及螺纹"的螺丝概念。但是这样就会缺乏连贯性。概念就是指我们的概念，是人类世界得以呈现的方式。螺丝是人类世界的一部分，只有在人类世界的背景中，它们才具有它们的身份。螺丝必然是根据我们对螺丝的概念向我们呈现的那种东西。也许自动螺丝刀处在一个不可与我们的世界通约的世界之中？如果是这样，我们试图去为那个世界表征认知就是毫无意义的。我们不能了解观察到青蛙的内容或黑色及螺纹的内容满足客观性制约的那个视角的意义。如果它们的世界不可与我们的世界通约，那么对于我们来说，它们就宛如没有世界（也见 Davidson 1974）。

地,一个有机体不拥有关于一个特性的概念*F*,除非它能对于它所拥有的对象的所有概念*a*,*b*,……认为*a是F*,*b是F*,等等。思维在本质上是结构式的①,因为世界在本质上是结构式的,反之亦然。

植根于客观性的形而上学的普遍性制约是与整体论制约相联系的,这一普遍性制约的巨大价值在于它为把概念运用系统与不能运用概念的系统分离开来提供了原则性的基础,对前者而言,存在着认知/世界的区别,对后者而言,则没有任何区别。以人类的听觉定位机制为例,信息加工机制根据声音速度和两耳之间的距离作出各种计算。我们可以说该系统表述了这些量。这一表述距离和速度的观念与我计划在斯坦福吃午饭时运用的表述有何类似之处呢?我们把概念看作是整个人的属性时,其正确性可以得到证明,那么,我们把概念看作是听觉机制的属性时,其正确性也能得到证明吗?

根据普遍性制约,如果定位机制也有认为*这一声音以y米/秒传播*及*耳朵间距离是x米*的能力,它才能认为*这一声音以x米/秒传播*(这里x米/秒是声速)和*耳朵间距离是y米*。但是*这声音以x米/秒传播*这个内容所具有的那种根据,*这声音以y米/秒传播*这个内容或*耳朵间距离是x米*这个内容并不具有。然而根据普遍性制约,对认为*这声音以x米/秒传播*或*耳朵间距离是y米*的能力来说,任何东西都不能成为正当根据,除非它也是认为*这声音以y米/秒传播*和*耳朵间距离是x米*的能力的正当根据。因而对概念*x米/秒*和*y米*来说,没有任何恰当的

① 对这一观念的讨论见坎贝尔(Campbell 1986)。

正当根据。所以虽然就广义的“表述”而言，可能存在着把表述归因于声音定位机制的理由，但是把概念归因于这一机制是不正确的。类似地，根据整体论制约，在一个系统尚不能想到(n - 1)的情况下，假定这一系统拥有某个数字 n 的概念，是毫无意义的。还有……

埃文斯(Evans 1982)运用普遍性制约证明了，拥有非概念知觉内容特有的信息连接[1]，并不能充分保证拥有概念。如果我对我的咖啡杯的记录仅仅在于当前我与杯子之间的知觉信息连接，那么虽然根据我的信息连接着传递关于杯子颜色的信息这一事实，我们能够设想把*那只杯子是灰色的*这一思想归因于我(参考*那只苍蝇在我的右边*)，然而却没有根据把*那只杯子是在斯托克制造的*或*在灯熄灭时那只杯子不会移动*或*那只杯子明天会打碎*这样的思想归因于我，即使假定我具有这些特性的概念也不行。其原因是，如果我记录杯子的能力因我(在信息连接的基础上)辨别杯子颜色和形状特性的能力而枯竭，那么我就不能把这只杯子记录为那种存在于黑暗中的，或在某一别的地点、某一别的时间制造的东西。此外，如果关于这个那只杯子的示范性思想仅仅在于我与杯子的信息连接，那么就不存在我曾对杯子有错误思想的基础。仅仅植根于与世界的信息连接中的认知观念，将是一个对世界的完善认知观念，因而也将是一个在认知与世界之间不作出任何区别的观

① 见埃文斯(Evans 1982:ch.5)。信息连接是有机体与对象之间的连接，有机体通过该连接接受关于对象的信息。一个人的判断和运动可能在信息连接的基础上对一个对象的特性改变作出响应。由有机体搜集起来的信息内容不是概念的。

念。(我与来自杯子的感觉传递之间的差距,不会大于,并且正好就是,我与我的视网膜传递之间的差距。但是后一种差距不是**认知**差距:我的认知在任何意义上都不是**关于**我的视网膜的。)

这些客观性制约为评估物理系统对概念的拥有提供了一个形而上学的可靠方法,正如自动螺丝刀、青蛙、听觉定位机制和与杯子的信息连接这些例子所表明的那样。但是,它们还有作为非概念论心理学事业成功判据(目标)的功能。非概念论解释是关于物理系统如何可能完成从处在只能利用适合于非概念内容的说明来表征的状态中到处在近似地满足客观性制约,因而也能用适合于概念内容的方式来说明的状态中的转变的解释。处在可用适合于概念内容的方式来说明的状态中,就是能满足某些逻辑规范:例如,这些状态,或由这些状态构成的复合体,能够进入正确的推理关系模式,能够接受真值评估,因而能承载起真正的、充满活力的意向性。要说明如何建立近似满足客观性制约的、以非概念方式表征的结构,就是要说明客观性如何能在物理世界中显现。

8. 视角依赖性

为当前探讨的客观性提供了一个原则性的基础,我们就能够解释为什么许多类型的 CTC(§4)没有把世界作为客观世界呈现给主体。同时我们也能够理解,必须把什么样的转换用于这些类型的 CTC,才能产生**确实**把世界作为客观世

界呈现给主体的 CTC 类型。

视角依赖能力与视角独立能力

概念内容表现为任务域经验中的可获得性，而 CTC 内容表现为基底域经验中的可获得性（§4）。非概念论的心理学解释取决于证明存在着一系列的 CTC 内容，在这些内容的一端，由于近似满足了客观性制约，基底域的经验可获得性势必要求任务域的经验可获得性。我想提出的是，按照经验上可获得的基底域能力的**视角依赖**程度，可使这个系列排成为一个尺度。据以对 CTC 内容作出典范说明的能力的视角依赖程度，就是内容不能满足客观性制约的程度。反之，在这些能力中取得视角独立性，是近似满足客观性制约所必需的，因而也是使内容呈现任务域所必需的。所以借助不同程度地依赖于视角的能力所说明的整个内容范围的理论，是只涉及这范围的一端的概念内容理论的概括。

有一种视角依赖能力是找到穿过城市的道路的能力，在这里，这种能力取决于一个从城市中的某个特定位置出发到达该城市的另一个特定位置的能力。出发位置是根据它的外貌确认的，然后，追寻从出发到结束的路线的能力将取决于对沿途的每个地点标志的识别，方法是部分根据它的外貌，部分根据它是从出发位置开始的一系列地点标志中的第 n 个地点标志这一事实。与每一地点标志相联系的是确定方向的指令："向右转，然后一直走"，等等。如果这个任务要求的不是从出发位置开始，或者从出发到结束不是沿着特定的路线行走，或者是走到一个不同的结束位置，

或者如果任何地点标志的外貌或相对位置改变了，那么这种能力就无法产生令人满意的表现。在所有这些情况下，这种能力深深地依赖于系统必须采取的关于区域的视角。这个区域里的某些(从出发位置和从地点标志得出的)**视点**处于优先地位，因为找到自己穿过城市的道路的能力依赖于占有这些特定的、优先的视点(也许按照一定顺序)。视点的观念是一个经验观念。在周围找到自己的道路的视角依赖能力是一种取决于系统具有某些特种经验的能力，系统因优先视点而享有那些经验。

找到穿过城市道路的视角独立的能力是这样的：如果一个人从一个人孔进入城市，那么无论他在哪里进入，他都能排除外部障碍物，找到去城中任一别的地点的道路。这种找到道路的能力不取决于一个特定的(从出发位置的)视角，或一组特定的视角(一条路线)。不存在优先视点，所以找到穿过城市的道路的能力，就是一个人**无论**在城市中具有**何种**视点，都能产生令人满意的表现的能力。这种能力当然依赖于经验，但是不依赖于系统具有的特种经验(地点标志 1 型经验之后跟着地点标志 2 型经验，等等)。无论系统碰巧享有何种经验视角，表现都能得到保持。

这里我们得出的一般观念是在一个区域中找到道路的基于经验的方式的观念，这些方式或多或少地取决于具有特种经验，即从优先视点得到的那些经验。这一真正的空间实例提供了最直接的例子，然而无论该区域多么抽象，我们都能得出一个人找到穿越区域的道路的视角依赖方式与视角独立方式之间的对比。例如，特性的识别可能取决于该特性得到满足的背景，无论这一特性是属于咖啡、电磁学或“自由战士”的。

从视角依赖能力到视角依赖内容

有了能力的视角依赖性与视角独立性之间的对比，我们就可以把这对比应用于参照这些能力典范地说明的非概念内容。因此，**视角依赖内容**是参照视角依赖能力典范地说明的内容。由于内容的内在性质，[1] 它只能从一个特定视角或限定的一组视角来考虑。

这是一个十分严格的关于内容的视角依赖性的观念，因为它是一个影响内容本身的观念，而不仅仅是影响那些使内容可被掌握的外部条件或使内容已被学到的条件的观念。所以我记录我母亲的能力是一个在许多方面取决于具有儿子的视角的能力。然而我还把她记录为某个可能具有许多别种关系的人；记录能力的这类视角依赖性（它在很大程度上与利用基底域概念说明 β 内容的理论无关）不影响这种能力的内容。我对一棵榆树的记录是从非植物学视角产生的记录（的确，处在我当前的状况，我只能从非植物学视角来掌握 * 榆树 * 这个内容），然而它是像从植物学视角也可获得的那种东西的所指对象的记录。[2] 如果我不是睡着了，或至少不是在做梦，我只能掌握 * 草是绿的 * 这个内容。但是内容本身（我所掌握的东西）的性质在很大程度上不受这个在内容以外的视角依赖性的影响。

因此，视角依赖内容是作为借助视角依赖基底域能力在内容理论内典范地说明的内容被引入的，与只涉及以可使内容被

---

① 在内容理论对于内容的认知意义的典范说明中获得的东西。

② 见帕特南（Putnam 1975）。

掌握为条件的依赖性的通常观念相比，视角依赖内容包含着一个十分不同并且彻底得多的视角依赖观念。不大彻底的观念对于解释某些状态的内容为什么不能满足客观性制约，是根本不起作用的。而比较彻底的观念则是这种解释的基础。

考察一个极端的例子也许是有帮助的。疼痛内容是通过处于疼痛中的能力来说明的。但是这种能力是大大地依赖于视角的：作为一种与世界打交道的方式，它取决于仅有的一种特定经验：经历疼痛，或回忆疼痛经验。[1] 这种能力取决于唯一的、特有的视点。这种视角依赖性确实影响着内容；疼痛经验不把世界呈现为与如何得到经验无关的方式，即不呈现为不能得到疼痛经验的运用概念的生物体可获得的某种方式。与之不同的是，视角无关内容则把世界呈现为是任何运用概念的生物体的视角都可潜在地获得的。[2] 所以如此的原因是，无论一个利用这些内容的生物体有时会用何种特定经验通过该区域，这些内容都会成功地发挥作用。

视角独立性、客观性制约和任务域

正是由于某些内容是接近于与视角无关的，所以才近似于满足普遍性制约。把我的咖啡杯呈现给我的那些内容，是把杯子作为那种对于例如斯托克的杯子制造商的经验是可

---

① 有趣的是，疼痛记忆是多么地不完全。虽然人们能以非概念方式回忆处于疼痛中的情况，但是关于疼痛的剧烈程度的记忆往往是很不可靠的。有人说要不是这样的话，世界上就只有独生子女了。

② 注意，这并不是说存在着不从特定视角认知世界的方式。每一个认知方式都是有视角的，但是只有某些认知方式把世界呈现为（近似地）是任何概念视角都可获得的东西。

获得的东西呈现给我的。所以我能接受我的咖啡杯是在斯托克制造的这一思想。

在内容不能满足普遍性制约的地方，这一内容就没有形成完全的结构，所以它就不能进入例如整个推理范围。我们可以用连字符来表述结构的缺失，使通常的概念内容标记法得到修正。举一个未必发生的例子，我可能有一个关于琼斯的记录，这个记录依赖于采用关于琼斯的“吃东西视角”（就像青蛙关于苍蝇的记录并不是一个完全空间的记录，而是一个“依赖于方向和运动的记录”一样）。如果把我的内容解释为一个概念的做法是正确的，那么我就可以认为*琼斯正在吃东西*。但是因为我的那个用来典范地说明我的琼斯内容的识别琼斯的能力依赖于吃东西视角，所以我就不能接受*琼斯用胶布封住他的嘴*或*琼斯从前是曼彻斯特的一个受精卵*的思想，即使我有*……用胶布封住他的嘴*和*……从前是曼彻斯特的一个受精卵*的概念。因此我的内容由于是依赖于视角的而不能满足普遍性制约；因而它是无结构的，应该表述为*琼斯-正在-吃东西*。

类似的情形还有，我把城市内远处的位置想成*那儿*的能力，是通过依赖于路线的在城市里找到我的道路的能力被典范地说明的。我处在这儿时可以想*那儿*，这时就把思维抛向我可以达到的目标位置之一。但是我接受关于“那儿”的思想的能力取决于我采用了“这儿”的视角（对特定的一组位置而言），因而是依赖于视角的，所以不能满足普遍性制约。我的内容实际是*这儿-到-那儿*。

因为某些内容是依赖于视角的，所以它们没有把它们的对象呈现为任何内容视角都能获得的那种事物，因此，它们不

满足普遍性制约。如果我们尝试以概念论方式来说明它们的内容,那么我们将会失败,因为这一内容不能满足普遍性制约,因而不能借助与客观的世界条目的语义关系被典范地说明(普遍性制约是对客观性的制约)。视角依赖性势必造成客观性制约的失败,继而势必造成典范的任务域说明的失败。它们没有把世界呈现为客观世界,正是因为某些非概念内容是依赖于视角的;它们成功地被算作把世界呈现为客观世界,正是因为某些非概念内容很少受到视角依赖性的影响。视角无关性势必造成对客观性制约的满足,继而势必造成典范的任务域说明。这表明,为了得到对概念能力的非概念论解释,所要求的那种心理计算建模是:将心理计算转换定义在具有减小内容的视角依赖性的效果的非概念内容之上。我即将着手研究怎样可能取得这种效果。

任务域的独立性和学习的中心地位

视角依赖性的一般情形可以简便地陈述为:整个系统或能力在心理学上的成功(而不是对一个能力在特定时间的特殊运用的评价)取决于就任务域而言(按 §4 中定义的意义)作出评价[①],而通过这些能力典范地说明的内容是依赖于视角的,因此不能作为拥有概念的根据。青蛙的“苍蝇思想”不是真正的苍蝇思想,因为它们的成功(因而它们的内容)取决于青蛙任务域的专有特征(成功地捕捉到苍蝇的益处,在重要性上超过了用舌头猛击众多远处目标的代价);青蛙的“认

① 或者按屈森斯(Cussins 1987 年 5 月)的理解,是处在任务域内。

知"依赖于特定任务域的视角。它不能进行概括。我的"琼斯认知"取决于吃东西任务域,我在帕洛阿尔托(地名——译者注)周围找到道路的能力,取决于帕洛阿尔托是一个有路线结构的任务域。(若在主干道十字路口重新设计几幢建筑物,我的整个能力就会被抹去。[①])

使一个系统的内容必然具有视角依赖性的方法,是使该系统的成功依赖于任务域。所以相反地,要使一个带有视角无关内容的系统的建立成为可能,人们必须将这系统建立得使它的成功不取决于某个任务域的偶发事件。出乎意料的是,要使一个认知系统成为一个具有可参照任务域典范地说明的内容状态的客观系统,以非概念方式表征的整个系统的能力必须是与任务域无关的。它必须在它有能力作自我修正以便在任务域之间转换的层次上运作;亦即在导致任务域的记录的那些学习过程的层次上运作。于是,这一非概念论理论的形成[②]就把学习放

---

① 刚搬家到帕洛阿尔托时,我习惯利用大街拐角处的TACO(墨西哥煎饼——译者注)的紫色招牌,从公路靠我这边的千篇一律的建筑物和道路中识别出我所住的EI CAMINO大街。一天,我驶过了我住的大街数英里。本地住宅公司已经对这种颜色提出异议,并要求TACO招牌的主人重新漆成灰色底色。

② 有些人认为情况恰好相反,即认为达到智能认知的路线是利用特定任务域的特征,有关的例子见伊思雷尔(Israel 1987),或巴怀斯(Barwise 1987),或罗森斯海因(Rosenschein 即将出版)。本段说明为什么学习是认知的核心,为什么"问题求解"(在GPS意义上)处在外围。重要的是要认识到,视角独立性像客观性制约一样,是没有一个物理系统能完全逼近的一种理想;我对帕洛阿尔托的认识能力一直处于路线不纯状态。问题关键是,这一理想确定了一个尺度,按照这个尺度,可对不同系统的概念性进行评估,从而可对它们的智能进行评估。使一个人的智能理论(如处境理论)把任务域依赖性变成一个优点(这样,用来设计成功系统的方法就是使系统处在任务域内),就是放弃心理学理论的形成,不管它可能是多么好的工程。一个理论可算作科学心理学理论,当且仅当它表明何谓改变一个物理系统使它更加近于概念系统。我对帕洛阿尔托的认识能力虽然是路线不纯的,却是一个概念能力,因为它在本质上是一个灵活的学习系统的一部分,这个系统带着经验,正使我的能力沿着具有越来越大的视角无关性的尺度移动。我进步得很慢,但过了最初两个月以后,我的确可设法不靠特征或地点标志,如TACO招牌上的颜色,去认出我住的大街了。

在认知的核心位置上,而不是放在我们为了达到认知本身所必须经过的一个外围转换阶段上。

总之,内容因其是依赖于视角的[①]而成为非概念的,所以我们能把这一从非概念内容到概念内容的转变表述为是从视角依赖性到视角独立性的转变。我需要的关于琼斯的记录,是从我能对琼斯采取的任何视角获得的记录。例如,我必须能够认出他,或知道何谓认出他,不仅在他吃东西的时候,而且在他游泳的时候,阅读的时候,在他用胶布封住自己嘴的时候,以及作为从前曼彻斯特的一个受精卵的时候。因而我认出琼斯的能力必须与任何特定的任务域无关。类似地,我必须有一个不依赖于路线的关于帕洛阿尔托的记录,使我能够例如从城市任何地方的人孔突然冒出而找到通往城市中任何别的可能是我的目的地的地点的道路。这不是没有任何视点(无论该视点是什么样的)的关于帕洛阿尔托的记录,而是不依赖于一个人在帕洛阿尔托内部占有某个特定地点(或一组地点,例如一条路线)的记录。我的关于帕洛阿尔托的记录是从任何地方进行的观察,因为无论我从何处冒出(也就没有预定路线)我都能找到我的道路。脱离视角依赖性,从而达到客观性,不是通过以某种方式使一个人的认知活动完全与视角(上帝眼睛的观察)相分离,而是通过使它越来越好地接近于无论他采取什么视角都能有效的方式。认知活动不能与视角相分离,但是它必须与任务域相分离。

---

① 更严格地说:"因其是借助依赖于视角的能力被典范地说明的"。凡用到时,这一扩展形式都应理解为较简短的表达形式。

9. 计算载体和认知地图

至此,我已经考察了通过整个系统的能力来说明的整体系统的内容。这些内容所用的**表述载体**是整体系统的能力。我已经指出了这些内容的非概念分析的形式。我已经表明,必须完成什么,才能使这种分析产生对客观性制约的近似满足,因而使表述载体携带具有构成性推理作用的概念内容。这些全都处于心理学解释的层次上(图 15-2)。最后,我们必须进入下意识层次去检验一下这些内容的**计算载体**,通过这些载体,认知科学将建立内容的模型。

地图绘制和地图使用

为了遵循 C3 的纲领,我们需要看一看用来降低系统表述状态内容的视角依赖性的计算方法。完成这一任务的有效研究策略,是检验具有这种效应的外部交际表述的转换。从 C3 的视点作这种考虑的重要性并不在于构造出来的交际表述(白板上的地图或标记),而是在于表述**构造**和表述**使用**的过程。C3 模型必须以计算方式来完成对这些过程的模拟。

我们考察一下普通地图是如何起作用的。我们可以设想,地图是由在一个地区各处行走的人构造的,他从他的每一视角观察该地区在他看来是什么样子。地图绘制者穿越这个地区,从而得到以自我为中心的关于整个布局的记录;即通过事物如何与**地图绘制者**在空间上发生关系来记录这些事物位

于何处。这是一种视角依赖能力，因为假如地图反映的仅仅是这一知识，它提供的关于该地区的记录将取决于沿着穿过这一地区的一条特定路线：地图制造者所沿的路线。所以对于希望沿一条穿过这地区的不同路线的任何人来说，它将是毫无用处的。我们可以说，这样的“地图”表述的只是自我中心空间，而不是客观空间；正如只懂得非概念内容的有机体只能自我中心地而不能客观地记录世界一样。

注意在这些早期阶段上，在地图使用者关于他在哪里的记录与他关于在这个位置有哪些特性的记录之间，并不存在无关性。这时，地点只是由它们看上去像什么或由它们与地点标记的相近性来确定，所以如果一个人发现自己是在一个没有树木的地点，那么他就不可能是在由简图表述标记为有树木的位置上。这样一个简单的地图不会弄错它在它表述的地点是什么样子，因为它表述构成这一空间（地点内容）的那些位置的方式，完全取决于它如何表述它在那些地点是什么样子。如果一个人发现自己在客观世界中的一个没有树木的地点，那么他就不处在由图形表述标记为有树木的位置。这一简单空间表述的认知模拟将会是这样的：它提供关于一个地点的记录，关于这个地点，人们只能认为它具有它看来具有的那些特性。例如，人们不可能认为*他们砍倒了这个地点的树木*。事物在人们认知中看来像什么样与事物是什么样之间没有区别。

随着时间的推移，地图绘制者可能沿着穿过这地区的许多不同路线行走，从而变得能够表述基于多条路线的关于这一地区的记录。对这知识的初步图形表述，可以表述一些特定路线，以及人们可用来确认自己是在这路线的什么地方或

是在哪一条路线上的那些特征。(很多旅游地图就是这样的。)假如某人也用这样的地图,他就将具有关于这个地区的**依赖于路线的**知识。地图变得越复杂,位置的确认就变得越加可能根据一个人曾经去过哪里和曾经如何旅行的知识,而不仅仅根据地图表述为这位置真正具有的那些特性。对认知模拟来说,存在着事物看来像什么样与事物是什么样之间的区别的萌芽,这是错误的可能性的萌芽。认知地图还不能确立关于表述位置的虚假判断的基础,但是可以应用另外一些不太精致的规范:例如,这个地图可能是骗人的。[1]

当一个人沿着穿过这空间的许多路线行走,并且使用他的跟踪技能时,他就能确定某些地区以不同视角来看是什么样的,不同视角是如何互相联系的,所以他能够着手绘制我们认为是该空间的(拓扑)地图的那种东西。这样的地图记录了该地区的空间状况,所以每个地点都以与所有别的地点相同的方式来表述[2](而不是在第一阶段根据它与地图绘制者位置的关系;或在第二阶段根据它与这一路线上另一些地点的关系)。任何一个地点或任何一组地点都不处于优先表述地

---

① 这里的思想是:与一系列越来越精致的非概念内容相关联的,有一系列越来越精致的、可应用于非概念内容的规范。真/假只能用于概念内容(因而只能用于其非概念内容充分地与视角无关、可近似满足客观性制约的那些状态)。但是也有一些可用于较低程度的非概念内容的较低的规范。在英国靠右行驶是错误的,靠左行驶是正确的。但是在接近本世纪初的年代,就没有关于沿着道路哪一边行驶是正确或错误的问题。由于交通量变得越来越大,靠左行驶的约定才开始建立起来。在这中间过渡时期,靠右行驶并不是错误(像现在那样),但是可能有危险。判断有危险的规范,比判断有错误的规范要低级一些,但毕竟是一个规范。

② 见埃文斯(Evans 1982:ch.6)。B·史密斯使用"从任何地方观看"这一短语,以示与内格尔"不从任何地方观看"的不同。

位。一个人一旦拥有这地图，他对这空间的记录就不限于在沿着穿过这地区的某些路线行走时感到该地区看起来是什么样的了，而是以非特定视点（从任何地方观看）的方式对这空间作出记录；**无论**一个人发现自己在这空间内的**什么地方**，该记录都具有实用价值，在这阶段，地点确认对特性确认的依赖，与特性确认对地点确认的依赖是相同的。由于地图表述的整体性，即使当地的采石场可能把一个山坡移走，人们仍然认得出这就是自己以前所在的地方。因此地图使用者能够说出，现在地图关于山坡的表述不对了。因而一个具有完全成熟的认知地图的主体就能形成关于位置的正确判断。他或她可能认为*这里曾经有一个山坡*，因为关于这个位置的思维方式不再因正在接受关于该地点现在看来是什么样的信息而消失。

这个描述通过地图使用者确认了他在周围找到自己的道路的渐少依赖于视角能力的序列，同时它又通过地图绘制者确认了越来越少依赖于视角的能力的渐进构造。此外，这个序列看来正是 C3 要求的那种序列，因为在构造更精致的地图时，那些可能的错误种类，因而客观性的层次，都变得更精致了。所以 C3 的任务就是说明如何为每一认知领域提供地图绘制和地图使用描述的计算模拟。

获得**认知**的视角独立性，可以认为是以计算方式构造**认知**的地图绘制和使用能力（O'Keefe and Nadel 1978）。在极端情况下，拥有认知地图势必要求以与所有别的地点相同的方式来看待每一地点。当然，任何实际的认知地图都不能完全达到这一目的，因为地图总是有界线的。但这正是应该存在的情况，因为客观性制约是一些理想化，原理上说任何事物都

不可能绝对逼近它们。然而关键问题在于，以自我中心的视角依赖能力为基础来构造认知地图，是达到对客观性制约的相对逼近的正确方式。如果主体在想到地点 A 时与他或她想到地点 B 时能采用相同的方式，那么持有任何关于地点 A 的思想（*A 是 F*）的能力，势必要求关于地点 B 想到同样思想（*B 是 F*）的能力。

我所说的地图绘制描述仅仅以时间上的位置记录为根据，这可能会令读者担心。我自行获得了特性记录和时间记录，但是若假定人们也能为这些记录作出类似的描述，似乎并不是太牵强的。我们谈的是我们在时间上自我定位的方式，所以关于“时间上的空间”的地图的思想似乎是相当自然的。在特性方面也有类似的情况：作出一个关于我们的辨别能力怎样才能变得越来越少依赖于视角的地图绘制描述似乎也是很自然的。此处非常需要大量的细节，但是我在本文中的宗旨始终是弄清 C3 怎样才能提供一个替代 LOT 的可能方案，而不是预言由 C3 提供信息的认知科学中若干年的详细经验研究的成果。

联结论：是一个适合于 C3 的计算构造体系吗？

联结论[①] 能提供一个适合于按照 C3 框架建立认知模型的计算构造体系吗？

---

① 见鲁梅哈特、麦克莱兰和 PDP 研究组（Rumelhart, McClelland and PDP Reseach Group 1986）。

斯莫伦斯基(P.Smolensky 1988 a)曾认为,恰当的"联结论处理方法"(PTC)——为了提供一个可供替代的建立认知模型的方式——要求联结论在"符号"与神经之间的层次上以"亚概念"方式建立模型。如果他是正确的,那么根据本文中的讨论,会立即得出这种看法:PTC不可以把给出概念论的心理学解释作为目的(§5)。概念论者不承认原子概念层次之下的任何表述层次;符号概念层次之下的任何解释层次都是实现方式的理论。但是PTC必须在概念层次与任何实现方式理论层次之间打开一个认知解释方面的缺口。所以PTC不能是概念论的。

通过图15-2(§1)的层级体系,向下追踪这一后果,就会得出:PTC需要利用一个不同于概念内容观念的内容观念,因而需要利用一个不同于S/S表述理论(§4)的表述理论。但是我们也看到(§3),只有运用了S/S理论,LOT方能说明计算心理学怎样能回答认知体现问题。

所以PTC看来是被捆住了手脚:以我们仅有的说明为基础,PTC要回答认知体现问题必须运用S/S理论。但是,为提供它所想望的认知替代方案,它不能把给出概念论的心理学解释作为目的。只要它不运用概念内容,因而也就是不运用S/S理论,它就只好寻找某个替代方案。正如福多尔和佩利舒(Fodor and Pylyshyn 1988)认为的那样,S/S理论是PTC的劲敌。

但是我希望,非概念内容的CTC观念如何使替代S/S理论的方案成为可能这个问题已经得到证明(§4, §6, §7, §8和§9),尽管S/S理论也能论及认知体现问题。C3能够做到这一点,因为(1)CTC和CTC的载体在解释上不是相互无关的,而解释无关性对于内容的语义说明与句法之间关系来说

是必不可少的;(2)C3 提供了一个不能被语义学理论说明的非概念内容观念,(3)存在着一个连贯的、非概念论的心理学解释观念,这个观念可以建立在 CTC 观念的顶上,同时通过这个观念,我们能够以非概念方式说明自然界中存在着运用概念的生物体。PTC 和 C3 像 GOFAI① 和 LOT 那样密切配合而结为一体。

但是我预料到会有这种观点:联结论构造体系适合于用 CTC 建立模型。与图 15-2 的层级体系中层次之间的大多数其他联结不同,这一联结在很大程度上是经验的。但是,读了大量的联结论文献,而没有强烈地感到需要一个非概念内容观念,是不可能的。我会接受某些简单的、仅仅是提示性的关于认为联结论构造体系适合于 C3 建模的理由。这些理由只能是提示性的,因为这是一个经验的问题,但无论如何,没有任何理由说冯·诺伊曼计算构造体系不能用于建立 C3 模型:②虽然在冯·诺伊曼构造体系中实现 S/S 理论是很自然的事,但是这些构造体系的力量远远超过了 S/S 理论所能对它们作出的利用。也许 PTC 对 C3 的需要超过了 C3 对 PTC 的需要。

斯莫伦斯基认为,要使联结论为思想的产生能力建立模型,它必须利用通过分布式表述获得的构成性表述结构。斯莫伦斯基(Smolensky 1987,1988 a)(用一个精心设计的玩具型例子)考察了联结论如何以结构方式表述 * 有咖啡的杯子 * ,

---

① 豪格兰的术语,表示"有效的老式人工智能"(Good Old Fashioned Artificial Intelligence)。

② 它仍是"C3":概念的计算构造。("概念的计算构造"的英文原文 The Computational Construction of Concepts 和"概念的联结论构造"的英文原文 The Connectionist Construction of Concepts 都包含 3 个以 C 开头的单词,所以都简称 C3。——译者)

而不是把它作为与*杯子*关联的词条和与*咖啡*关联的词条之间的句法关系。他的例子有助于实现我们的目标,因为它指出了联结论自然地适合于C3建模的第一个理由:联结论系统自然地采用依赖于视角的表述。图15-3出示*有咖啡的杯子*的联结论表述的表示法。

单元	微特征
●	立式容器
●	热液体
○	接触木制品的玻璃杯
●	瓷器的弧形表面
●	烧焦的气味
●	接触瓷器的褐色液体
○	银白色椭圆物体
●	一指长的柄
●	带弧形边和底的褐色液体

图15-3

图4是*没有咖啡的杯子*的联结论表述的表示法。

单元	微特征
●	立式容器
○	热液体
○	接触木制品的玻璃杯
●	瓷器的弧形表面
○	烧焦的气味
○	接触瓷器的褐色液体
○	银白色椭圆物体
●	一指长的柄
○	带弧形边和底的褐色液体

图15-4

为了了解在联结论系统中*咖啡*的表述可能是什么,我们只须从*有咖啡的杯子*的表述中减去*没有咖啡的杯子*的表述。其结果的表示法见图15-5。

单元	微特征
○	立式容器
●	热液体
○	接触木制品的玻璃杯
○	瓷器的弧形表面
●	烧焦的气味
●	接触瓷器的褐色液体
○	银白色椭圆物体
○	一指长的柄
●	带弧形边和底的褐色液体

图 15-5

这样做的目的在于看清*咖啡*的联结论表述是深深地依赖于背景的,它是*处在—杯子—背景—中的—咖啡*的表述。于是还会有这些表述:*容器—中的—速溶—咖啡—颗粒—背景—中的—咖啡*,*全都—洒到—我的—纸—上的—咖啡*,*处在—咖啡—喝得—太—多—看起来—有点—苍白—的—某人—背景—中的—咖啡*,等等。显然,对*咖啡*的这些依赖于背景的表述中的任何一个进行组合的推理可能性,将受到在这背景中适合的推理的限制。所有的咖啡都有助于你的消化。但是"处在—洒到—我的—纸—上—的—背景—中的—咖啡"具有相反的作用。这一*咖啡*概念起着很不同的作用。例如,任何有关咖啡能被说出的事情,都能用表达"咖啡"概念的单词来说出。但是为雀巢服务的广告商不能通过*处在—喝得—太多—看起来—有点—苍白—的—某人—背景—中的—咖啡使你觉得好极了*的表达来为雀巢咖啡作广告。

因此,联结论表述自然地是依赖于视角的,所以不能满足普遍性制约。但只是对于输入和输出单元上的表述才需要如此。在形成联结论算法的过程中,大部分注意力指向了对于

怎样能从输入和输出表述是依赖于视角的系统中得到有用行为的理解。如果 C3 是正确的,那么这正是我们为了建立认知模型而想要的。隐蔽单元之间的加权值会逐步形成,从而使得包含这些单元的加工过程变得可以响应许多不同的依赖于视角的咖啡表述之间的联结。在已经论证的我们拥有 * 咖啡 * 概念必须——像所有概念一样——依赖于从大量视角依赖的咖啡表述中得出的简化视角的构造的背景中,这一点显然是令人鼓舞的。

认为联结论适合于 C3 建模的第二个理由是:与冯·诺伊曼的构造体系不同,"学习"是联结论的中心。这就较容易使人想到怎样在联结论构造体系中实现 C3 模型,因为 C3 把学习当做认知的中心,甚至当做认知的本质。

第三个理由是,在任何非玩具型例子中,联结论表述只能通过对输入模式与输出模式之间复杂联结的分析来解释。至于能够给出何种对隐蔽单元活动的意义的分析,常常显得有些神秘,特别是在输入单元直接与某些感觉系统相联结,而输出单元直接与某些效应器系统相联结的场合。① 正是由于未能看破这种神秘性,致使福多尔和佩利舒(Fodor and Pylyshyn 1988)在对联结论的批判中误入歧途。他们假定,微特征的"亚概念"空间必须被看作是对切成薄片的有效老式语义特征的量值分配:

① 有时有这种看法:这个难点只限于个别隐蔽单元的意义,而不是许多单元上的活动矢量。其理由是,根据假定,虽然隐蔽单元可能确实没有任何概念意义,但活动矢量将会有。我希望 C3 能证明这一点也并非是必然的。对认知所作的心理学分析是彻头彻尾非概念的。描述的概念层次只不过是以心理计算方式提供了对心理计算因果活动过程的无效约束。

很多联结论者认为,对应于常识性概念(**椅子、约翰、杯子**等等)的心理表象,“分布”在几批本身具有表述内容的较低层次单元上。使用普通的联结论术语,较高的或“概念层次”的单元对应于“亚概念”微特征空间中的一些矢量。**这里的模型是某种像被定义的表达式与它的定义特征分析之间的关系那样的东西**:所以概念**单身汉**可以被认为对应于一个特征空间中的一个矢量,这空间包含**成年人、人、男性**和**已婚的**;即被认为以正值分配给前三个特征,以负值分配给最后一个特征……。由于微特征常常被假定是从刺激样本的统计特性自动(即经由学习过程)得出的,所以我们可以认为它们表达的那些种类的特性是由对多组刺激的多元分析揭示的。特别是,它们无需与英语单词相对应;对于非专业人员需要用单词来形容的那些术语来说,它们可以是更为细小的颗粒,或者说是非典型的术语。**然而与此不同的是,它们纯粹是寻常语义特征,非常像词典编纂者惯用的表述单词意义的那些特征**(黑体字是我强调的)。

但这其实是由福多尔和佩利舒的偏见所致。我们可称之为“概念论的盲点”:内容的表述结构只能是它的概念结构,因为这里仅有的那种内容是概念内容。但是我不仅证明了情况并非如此,而且也证明了它不能如此。C3 提供了非概念内容的观念,并且表明了如何利用认知地图的构成中所包含的那些减小视角依赖性的变换公正地对待对认知的经典制约。联结论可能不是句法语义的,但是 PTC 是依据这一事实证明联结论怎样才能为建立认知模型的经典方式提供替换方案的。

反对联结论的观点常常以这种形式出现:一个黑箱,甚至

是一个十分成功的黑箱,并不为心理学理解增添任何东西(见§1)。在某个刺激环境中训练出来一个联结论网络,它在那个环境中的表现可能会十分成功,但是因为该网络已经通过它的学习规则系统自行编程,对这种表现是如何完成的,我们将一无所知,除非我们能够检测出它的隐蔽单元上的加权值,并**找到对这些加权值的解释,以便我们能理解为什么该系统是成功的**。经典人工智能也面临着类似的问题:处理 LISP 基本表达式的一个复杂模式怎样才能构成一个心理学理论? LOT 解决了这个问题,但是它提供的解答对联结论来说是无用的,因为通常任何连贯的语义学都不能定义在隐蔽单元之上。情况可能是这样的:关于单元联结性模式的功能发挥,所能确定的只是它协调某些输入活动矢量种类与某些输出活动矢量种类之间的联结的那种方式。那怎样能提高我们的心理学理解呢?

像联结性的模式一样,CTC 是根据它在输入与输出之间的协调力量被说明的。如果能表明由这些力量完成的过程怎样才能近似地满足对客观性的制约,这些力量就是意向性的力量。联结论理论家能够向我们表明,隐蔽单元的力量怎样能在训练过程中逐渐发展以构成认知地图吗? 如果能做到这一点,我们就能够以经验方式探索一种新方法,用以理解世界中的事物如何能对世界进行思考。

一个推测性的结论:认知的联结论载体

在§9 的第一部分中,我们看到了为逐步增长的视角的独立的能力所要求的、因而也是为变得逐步逼近客观性制

约的 CTC 内容所要求的那些种类的地图绘制和地图使用的计算过程。在§9的第二部分,我们看到了认为联结论提供了一个用来完成这些计算过程的适当的构造体系的一些理由,虽然毫无疑问,在一个很长的时期内,这仍将是一个悬而未决的问题。但是即使如此,我们仍可以就 CTC 和认知地图的联结论载体提出问题。

CTC 内容的**表述**载体是整体系统能力,但是这些内容也具有**计算载体**,这些计算载体是整个系统能力的特殊执行方式的下意识的因果基础。自然可以确认两种联结论载体:分布在许多加工单元上的活动模式,和许多加工单元之间的加权联结性的模式。前一些模式比后一些模式更具瞬态性,但两者都是以因果方式活动的:系统在下一瞬时的加工状态,是活动模式和加权联结性模式两者的函数。一个活动模式本身不会明确表现出任何特定的视角依赖性程度,因为它究竟是不是视角依赖**能力**的部分基础,将取决于导致该活动的输入的背景,它将导致的输出的背景,和它所属的那个系统的进一步的配置情况。然而我们能够在引申的意义上谈论某些隐蔽单元上的活动模式的视角依赖性,靠的是该活动模式在依赖于背景的输入表述与依赖于背景的输出表述之间所起的协调作用(如咖啡例子中)。在这意义上,十分精致的联结论系统和每一个粗略的联结论系统的活动模式将同样地依赖于视角。由精致的和粗略的网络实现的内容间的差异的出现,靠的是另一种联结论载体:联结性模式。

我们可以试着说,如果一个联结性模式以地图般的方式在依赖于背景的输入模式与依赖于背景的输出模式之间起

协调作用,那么它就完成了一个认知地图。地图绘制者必须对有关地域始终采用自我中心的、依赖于背景的表现方式作为他的地图制造的基础,而地图使用者为了在这一地域内行走,必须始终从这地图产生出一些自我中心的、依赖于背景的指令。因此,我们可以认为地图是一个从自我中心的、依赖于背景的表述到自我中心的、依赖于背景的表述的函数。从非特定视点表述测绘空间的精致地图,将属于使地图使用者能以与视角无关的方式在这个地域中找到他或她的道路的那一类功能。如果联结论网络内的一个联结性模式完成了从分布在输入单元上的自我中心的、依赖于背景的活动模式到分布在输出单元上的自我中心的、依赖于背景的活动模式的函数,那么它就能完成一个认知地图。如果这个函数属于能使与视角无关的能力顺利通过测绘区域的那类函数,那么这联结性模式就将完成一个精致的、植根于客观性的认知地图。

拥有正确类别的认知地图,是系统具有在被映射区域中找到它的道路的与视角无关的能力的因果基础。因为我们看到,在周围找到自己道路的能力所具有的视角独立性,是通过参照这些能力被典范地说明的那些内容而满足客观性制约的基础,所以把认知地图确认为系统拥有一个概念的因果基础,是自然的。因此我们可以说,概念是作为加权联结性模式在联结论系统中实现的。

一组给定的加权联结可以实现不止一个的认知地图。既然有了概念的整体论,我们的确本该预料到这一点。(未曾拥有一组以概念方式相联结的概念,却拥有一个概念,这在逻辑上是不可能的,这种不可能性的因果基础也许在于这一事实:

正是同一组联结,或几组广泛地交叠的联结,实现了对应于这些概念中每一概念的认知地图。)在极端情形中,系统所拥有的每一个概念,将在整个系统的每一个联结中实现。C3 框架的科学层次将使概念的拥有在因果上成为合法的,但是内容的概念表征在对系统行为所作的科学心理学解释中不会起任何作用。[①]

谨以此文纪念我的父亲 M·屈森斯(1905—1987)。

我要感谢下列诸位,对本文及对文中思想的形成所作的帮助:

D·布罗茨基,J·坎贝尔,D·查尔斯,B·蔡尔德,R·克里斯利,A·克拉克,M·达米特,J·豪奇兰,D·库尔曼,D·莱维,M·马丁,G·农贝格,G·奥布赖恩,C·皮科克,J·斯拉格,P·斯科科斯基,S·斯托尔内塔,B·史密斯,S·斯塔基和 D·塔塔尔。我受益于在下述地点的交谈:坦普尔大学,1988 年北卡罗来纳州哲学和心理学学会年会,学习问题研究所,施乐 PARC 的系统科学实验室,斯坦福 CSLI,在 D·查尔斯的奥丽尔讨论小组,伯克贝克联结论讨论小组。我深深感谢牛津新学院的青年研究基金、斯坦福 CSLI 的博士后基金给予的资助,感谢施乐 PARC 系统科学实验室的资助和 PARC 的优越的研究环境。尤其感激查里斯的鼓励和信任。

① 所以如果这一说明提供了一种意义,在这意义上概念具有科学实在性,那么这一意义是与概念内容的相当大的不确定性完全相容的。

参考书目

Armstrong, D. M. (1968). *A Materialist Theory of Mind.* London: Routledge & Kegan Paul.

Barwise, J. (1987). 'Unburdening the Language of Thought.' In *Two Replies*, CSLI Report No. CSLI-87-74.

Bennett, J. (1976). *Linguistic Behaviour.* Cambridge: Cambridge University Press.

Block, N. (ed.) (1980). *Readings in Philosophy of Psychology*, Vol. I. London: Methuen.

Brachman, R. J., and Levesque, H. J. (1985). *Readings in Knowledge Representation.* Los Altos: Morgan Kaufmann Publishers.

Campbell, J. (1986). 'Conceptual Structure.' In C. Travis (ed.), *Meaning and Interpretation*, pp. 159–74. Oxford: Basil Blackwell.

Churchland, P. M. (1979). 'Eliminative Materialism and the Propositional Attitudes.' *Journal of Philosophy* 78: 67–90.

Churchland, P. S. (1986). *Neurophilosophy: Toward a Unified Science of the Mind–Brain.* Cambridge, Mass.: MIT Press/Bradford Books.

Cussins, A. (Jan. 1987). 'Varieties of Psychologism.' *Synthese* 70: 123–54.

—— (May 1987). 'Being Situated Versus Being Embedded.' Stanford University, *CSLI Monthly* 2 (7):14–20.

—— (1988). 'Dennett's Realisation Theory of the Relation between Folk and Scientific Psychology', Commentary on Dennett: *The Intentional Stance. Behavioural and Brain Sciences* 11 (3): 508–9.

—— (1990). 'The Explanatory Role of Nonconceptual Content.' In D. Charles and K. Lennon (eds.), *Reductionism and Antireductionism.* Oxford: Oxford University Press.

—— (forthcoming). 'The Emergence of Objectivity: Why People are not just Complex Frogs.'

Davidson, D. (1967). 'Truth and Meaning.' In Davidson (1984).

—— (1974). 'On the Very Idea of a Conceptual Scheme.' *Proc. & Addresses of Amer. Philos. Assoc.* 47, and in Davidson (1984).

—— (1984). *Inquiries into Truth and Interpretation.* Oxford: Clarendon Press.

Dennett, D. C. (1969). *Content and Consciousness.* London: Routledge & Kegan Paul.

—— (1978). 'Toward a Cognitive Theory of Consciousness.' In D. C. Dennett, *Brainstorms: Philosophical Essays on Mind and Psychology*, pp. 149–73. Montgomery, Vt: Bradford Books.

—— (1987). *The Intentional Stance.* Cambridge, Mass.: MIT Press.

Dretske. F. (1981). *Knowledge and the Flow of Information.* Cambridge, Mass.: MIT Press.

Dummett, M. (1975). 'What is a Theory of Meaning.' In S. Guttenplan (ed.), *Mind and Language*, pp. 97–138. Oxford: Clarendon Press.

—— (1976). 'What is a Theory of Meaning (II).' In G. Evans and J. McDowell (eds.), *Truth and Meaning*, Oxford: Oxford University Press.

—— (1978). *Truth and Other Enigmas.* London: Duckworth.

Evans, G. (1982). *The Varieties of Reference*. Oxford: Oxford University Press.
—— (1980, 1985). 'Things Without the Mind—A Commentary Upon Chapter Two of Strawson's *Individuals*.' In Evans (1985), *Collected Papers*, pp. 249–90. Oxford: Clarendon Press.
Fodor, J. A. (1976). *The Language of Thought*. Sussex: Harvester Press.
—— (1980). 'On the Impossibility of Acquiring More Powerful Structures', and 'Reply to Putnam.' In M. Piatelli-Palmarini (ed.), *Language and Learning: The Debate between Jean Piaget and Noam Chomsky*, pp. 142–62 and 325–34. London: Routledge & Kegan Paul.
—— (1981*a*). 'Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology.' In Fodor (1981*b*).
—— (1981*b*). *Representations*. Cambridge, Mass.: MIT Press.
—— (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, Mass.: MIT Press/Bradford Books.
—— (1986). 'Why Paramoecia Don't Have Mental Representations.' *Midwest Studies in Philosophy* 10: 3–23.
—— (1987). *Psychosemantics*. Cambridge, Mass.: MIT Press.
—— and Pylyshyn, Z. W. (1988). 'Connectionism and Cognitive Architecture.' *Cognition* 28: 3–71.
Frege, G. (1891). 'On Sense and Meaning.' In P. Geach and M. Black (1960). *Translations from the Philosophical Writings of Gottlob Frege*, pp. 56–78 Oxford: Oxford University Press.
—— (1918, 1977). 'Thoughts.' In P. T. Geach (ed.), *Logical Investigations*, pp. 1–30. Oxford: Basil Blackwell.
Gibson, J. J. (1979). *An Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.
Goodman, N. (1951). *The Structure of Appearance*. Cambridge, Mass.: Harvard University Press.
Hofstadter, D. R. (1985). 'Waking Up from the Boolean Dream; Or, Subcognition as Computation.' In D. R. Hofstadter, *Metamagical Themas: Questing for the Essence of Mind and Pattern*, pp. 631–65. New York: Basic Books.
Israel, D. (1987). 'The Role of Propositional Objects of Belief.' CSLI report No. CSLI-87-72.
Lenat, D. B., and Feigenbaum, E. A. (1987). 'On the Thresholds of Knowledge.' MCC-AI Non-Proprietary Technical Report.
Lewis, D. (1966). 'An Argument for the Identity Theory.' *Australasian Journal of Philosophy* 63: 17–25.
Marr, D. (1977). 'Artificial Intelligence—A Personal View.' *Artificial Intelligence* 9: 37–48, and reprinted in this volume.
—— (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco: W. H. Freeman.
McDowell, J. (1977). 'On the Sense and Reference of a Proper Name,' *Mind* 86: 151–85.
Millikan, R. (1984). *Language, Thought and Other Biological Categories*. Cambridge, Mass.: MIT Press/Bradford Books.
O'Keefe, J., and Nadel, L. (1978). *The Hippocampus as a Cognitive Map*. Oxford: Oxford University Press.
Peacocke, C. (1986). *Thoughts: An Essay on Content*. Oxford: Basil Blackwell.
—— (1989). Inaugural Lecture, given in the Examination Schools, University of Oxford, during Trinity Term.
—— (forthcoming). 'Perceptual Content.' In a volume in honour of D. Kaplan, edited by J. Perry, J. Almog, and H. Wettstein.
Pellionisz, A., and Llinas, R. (1979). 'Brain Modelling by Tensor Network Theory and Computer Simulation.' *Neuroscience* 4: 323–48.

—— (1980). 'Tensorial Approach to the Geometry of Brain Function.' *Neuroscience* 5: 1125–36.
—— (1982). 'Space-Time Representation in the Brain.' *Neuroscience* 7: 2949–70.
—— (1985). 'Tensor Network Theory of the Metaorganization of Functional Geometries in the Central Nervous System.' *Neuroscience* 16: 245–73.
Perry, J. (1979). 'The Problem of the Essential Indexical', *Nous* 13: 3–21.
Place, U. T. (1970). 'Is Consciousness a Brain Process?' In C. Borst (ed.), *The Mind Brain Identity Theory*, pp. 42–51. London: Macmillan.
Putnam, H. (1973). 'Reductionism and the Nature of Psychology.' *Cognition* 2: 131–46.
—— (1975). 'The Meaning of Meaning.' In H. Putnam, *Mind, Language and Reality*, pp. 215–71. Cambridge: Cambridge University Press.
Pylyshyn, Z. W. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science*. Cambridge, Mass.: MIT Press.
Quine, W. (1960). *Word and Object*. Cambridge, Mass.: MIT Press.
Reifel, S. (1987). 'The SRI Mobile Robot Testbed: A Preliminary Report.' technical note 413, SRI International, Menlo Park, Calif.
Rosenschein, S. (forthcoming). *An Introduction to Situated Automata*. Forthcoming in the CSLI Lecture Note Series, Chicago Press.
Rumelhart, D., McClelland, J., and the PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1 and 2. Cambridge, Mass.: MIT Press/Bradford Books.
Schiffer, S. (1987). *Remnants of Meaning*. Cambridge, Mass.: MIT Press/Bradford Books.
Smart, J. (1970). 'Sensations and Brain Processes.' In C. Borst (ed.), *The Mind Brain Identity Theory*, pp. 52–66. Place: Macmillan.
Smith, B. (1987). *The Correspondence Continuum*. Center for the Study of Language and Information, Report No. CSLI-87-71.
Smolensky, P. (1987). 'The Constituent Structure of Connectionist Mental States: A Reply to Fodor and Pylyshyn.' *The Southern Journal of Philosophy* 26 Suppl.: 137–60.
—— (1988*a*). 'On the Proper Treatment of Connectionism.' *Behavioral and Brain Sciences* 11: 1–74.
—— (1988*b*). Department of Computer Science, University of Colorado at Boulder, Technical Report on Tensor Representation.
Stich, S. (1983). *From Folk Psychology to Cognitive Science: The Case Against Belief*. Cambridge, Mass.: MIT Press.
Strawson, P. (1959). *Individuals*. London: Methuen.
Swinburne, R. (1986). *The Evolution of the Soul*. Oxford: Clarendon Press.
Williams, S. G. 'Computers, Validity and the Runabout Inference Ticket', Worcester College, Oxford University unpublished paper.
Winograd, T. (1973). 'A Procedural Model of Language Understanding.' In R. C. Schank and K. M. Colby (eds.), *Computer Models of Thought and Language*, pp. 152–86. San Francisco: W. H. Freeman.